NOTICE: When government or other drawings, specifications or other data are used for any purpose other than in connection with a definitely related government procurement operation, the U. S. Government thereby incurs no responsibility, nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use or sell any patented invention that may in any way be related thereto.

AFCRL-63-316

64-3

EFFICIENT UTILIZATION OF CHANNEL CAPACITY FOR SPEECH COMMUNICATION

64-3

R. L. Brueck
A. R. Aitken
D. R. Ziemer

TEXAS INSTRUMENTS INCORPORATED
6000 Lemmon Avenue
P. O. Box 6015
Dallas 22, Texas

FINAL REPORT

PART II

Contract No. AF 19(628)-345
15-73801-15
Project No. 4610
Task No. 461002

July 1963

Prepared for

AIR FORCE CAMBRIDGE RESEARCH LABORATORIES
OFFICE OF AEROSPACE RESEARCH
UNITED STATES AIR FORCE
BEDFORD, MASSACHUSETTS

AFCRL-63-316

EFFICIENT UTILIZATION OF CHANNEL CAPACITY

FOR

SPEECH COMMUNICATION

R. L. Brueck
A. R. Aitken
D. R. Ziemer

TEXAS INSTRUMENTS INCORPORATED
6000 Lemmon Avenue
P. O. Box 6015
Dallas 22, Texas

FINAL REPORT

Contract No. AF 19(628)-345
15-73801-15
Project No. 4610
Task No. 461002

July 1963

Prepared for

AIR FORCE CAMBRIDGE RESEARCH LABORATORIES
OFFICE OF AEROSPACE RESEARCH
UNITED STATES AIR FORCE
BEDFORD, MASSACHUSETTS

ABSTRACT

This report describes a research investigation directed toward more efficient utilization of channel capacity for speech communication. This objective was pursued by a program (1) to theoretically analyze the benefits realized by "predictive encoding" of vocoded speech sources, (2) to propose and analytically design a model of a typical processing system utilizing predictive coding, and (3) to evaluate the performance characteristics of such a system by simulation with vocoded speech samples on a digital computer.

The results were significant in that a compression of 35 to 40 percent was obtained relative to the initial requirements for transmission of the spectrum portion of vocoded speech data. This magnitude of compression can be obtained for essentially real-time transmission and buffer storage requirements of the order of 100 bits.

Compression factors greater than one-half are possible if a time delay in the speech transmission can be tolerated and if additional memory can be supplied.

FOREWORD

The basic philosophy behind the research reported in this document grew out of an in-house program at Texas Instruments in imagery enhancement and bandwidth compression over the two year period 1958 to 1959. Beginning in 1960, these results were applied to the general problem of speech bandwidth compression with the intent to implement some of the advanced coding concepts of modern information theory into the speech bandwidth compression problem. Early in 1961 our preliminary results were discussed with the Air Force Cambridge Laboratories leading to the study contract for which this document is Part II of the final report.

Actual work on the contract began early in 1962 and continued until June of 1963. The initial study efforts were performed by Mr. A. R. Aitken who extended the predictive encoding concept beyond what was originally proposed. Mr. R. L. Brueck joined the program and took over the research efforts after Mr. Aitken left Texas Instruments to attend the University of Texas. Mr. Brueck was able to achieve some rather elegant solutions to what turned out to be formidable mathematical and statistical problems. Without the availability of some advanced computational algorithms, which he designed, it would have been difficult indeed to have completed the program as successfully as has been the case. The writing of the final report was almost entirely his activity.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF ILLUSTRATIONS

LIST OF ILLUSTRATIONS (Continued)

LIST OF TABLES

EFFICIENT UTILIZATION OF CHANNEL CAPACITY FOR SPEECH COMMUNICATJCATION

by

R. L. Brueck
A. R. Aitken
D. R. Ziemer

SECTION I

INTRODUCTION

Stated most simply, this research report describes an investigation directed toward more efficient utilization of channel capacity for speech communication. This objective was pursued by a program (1) to theoretically analyze benefits that might be realized by "predictive coding" of vocoded speech sources, (2) to propose and analytically design a model of a typical processing system utilizing predictive coding, and (3) to evaluate the performance characteristics of such a system by simulation with vocoded speech samples on a digital computer.

In an attempt to provide an interested reader with sufficient (but minimal) motivation and background, a brief discussion of fundamental concepts from information and statistical prediction theory is included as an introduction. These elementary remarks in no way constitute the whole of the theory upon which this research is based. The first three references* more than adequately provide the interested reader with the rigour and insight of the abstract theory. The sole interest here is to provide argument for "why" and "how" to pursue the insight of that theory, and to augment it where necessary. In particular, to proceed one must understand the necessity and manner of assigning measures to any coding system and clarify parameter trade-offs that are involved in processing a message from a source through a communication channel to a remote sink.

This study is strictly preliminary in extent. Expenditures of effort such as: the organizing, formulating, and sorting out of the relevant theory; the development of systematic and efficient means to accumulate and handle necessary source data; the investigation and solution of problems of accurate mathematical computation; the preparation of computer programs that are necessary to gather design and evaluation measures; and the unproductive results of chasing false or immaterial problems; are all pre-requisite to a more exhaustive program of investigation. As is to be expected, the return on

*See Section VI., References.

these efforts, as measured by insight and detailed conclusions, is weighted heavily toward the continuation of research. However significant the insight and conclusions of this work, more questions and new ideas have been generated than have been analyzed in conclusive detail. The concepts, investigative approach, and analytical methods employed in this program will be generalized and improved in future efforts. Even more certainly, additional research will augment, clarify, and possibly correct conclusions reached at this time. For these reasons, it is important to evaluate the extent to which the theoretical and analytical tools used actually predicted and "measured" (in the sense of reliable estimates) the outcome of experimental results. It is significant to establish this program, and those to follow, as formulated on a sound and useful analytical basis, and not solely on naked intuition or pure art of experience.

A. SOURCE PROCESSING CONSIDERATIONS.

The purpose of any communications channel is to transmit messages as generated by some source to some physically remote sink. It is intuitively clear that in some temporarily undefined sense, various information sources produce message "bulk" at different rates. For instance, total message bulk produced by a manually operated telegraph key in one second is much less than that being produced by a television camera in one second. Similarly, it is also clear that various communication channels are of different sizes in the sense that each may convey only some maximum message bulk per unit time, and in general, the larger a channel is the more expensive it is to build and operate. Finally, it is clear that the greater the rate of production of a given source the greater must be the size and cost of the channel provided to communicate its messages. However, for one or more of several possible reasons, it is not always necessary to transmit the entire message bulk through the channel. Therefore, an economical benefit can be realized if, prior to transmission, the input message is subjected to some transformation which reduces or compresses its bulk to that of the essential part and thus reduces the required size of the communication channel. Transformations of this sort are defined to be source processing operations.

Most frequently, when one talks of compression of a signal (the physical representation of a message), one means or at least implies, reducing its bandwidth. Implicitly then, bandwidth is employed as a measure of source and channel sizes. While bandwidth is certainly a very meaningful parameter of sources and channels, equating source size and required channel size on this basis alone turns out to be quite meaningless; ingenious modulation schemes exist whereby the channel bandwidth required to transmit messages from a source of given bandwidth can be made as large or as small as one likes by adjusting other system parameters

Information theory, as introduced principally by C. E. Shannon[1], provides alternate measures of source rate and channel size. Specifically, it is necessary to assign measure to: (1) the bulk or volume of a message by its information content; (2) the bulk rate (average rate of bulk production) of a source by its entropy, H, which is its average rate of information production; and (3) the bulk rate of a noiseless channel by its capacity, C, which is the maximum average rate at which it can convey information from source to sink. The basic significance of these measures is primarily in the fact that there exists a very meaningful relationship between the entropy of any source and the (minimum) channel capacity required to communicate its messages. Specifically, any channel is large enough to satisfactorily communicate messages from an arbitrary source of entropy, H, if it has a capacity $C \geq H$. However, if the channel is to be as small as possible (i. e., if the equality $C = H$ is to hold, or nearly so), the input messages must be transformed in a certain way prior to transmission. These transformations are the essence of efficiently utilizing channel capacity, and together with related transformations at the receiving end, are called source processing. In contrast, if optimum processing is not done, the channel must have some capacity greater than the source entropy.

Clearly then, a source processing transformation may be considered a compression operation, at least in the sense that required channel capacity is being minimized. However, if it happens that the specified channel has greater bandwidth than the specified source, then an optimum processing transformation may actually "expand" the input signal bandwidth. In such an instance, channel efficiency is maximized by adjusting other system parameters (such as transmitter power) which also figure in the measure of capacity. One can avoid this difficulty in pinpointing the meaning and description of efficiency, by describing a new reference for processing operations.

To resolve the apparent difference is purpose between channel coding in the usual context, and bandwidth compression operations on a source, one makes a clarification of our reference to channel and a distrinction in types of coding.

The over all coding transformation, (and also the decoding transformation at the receiving end), can be conveniently separated into two operations in series. These are called source processing and channel coding, respectively. Figure 1 illustrates this discussion. The principal feature is characterized by a definition of a "standard" or reference channel. The objective of source coding is to match the source to this standard channel input (considered as a "standard source"), to the actual channel. For present purposes, the importance of the standard channel lies in that it is defined so as to be minimum bandwidth channel in the same sense one ordinarily uses in discussing bandwidth compression. As a result, optimum source processing implies maximum compression of both bandwidth and capacity, and the distinction in objective is unnecessary. The significant measure of the standard channel is the information rate, it being understood that a trade-off exists through channel coding between this rate and the power utilized in the actual channel. Also, the error rate or fidelity of the actual channel is for the most part a separate concern from the fidelity of source processing operations directed toward reducing the information rate input to the standard channel.

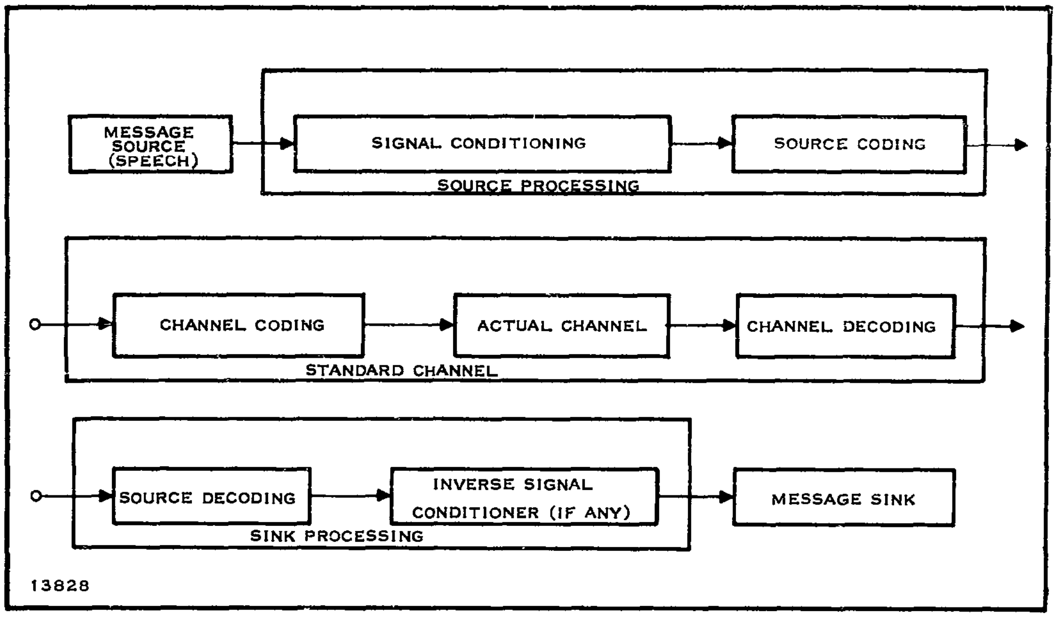


Figure 1. Speech Communication System—Simplified Block Diagram

The standard channel is assumed to be errorless. Thus, fidelity is to be interpreted in the sense that any operation on the source signal at the transmitter must be accomplished with a level of faithfulness of source reproduction at the receiver. Increased efficiency resulting from optimally matching source information rates to the standard channel is only meaningful if measured subject to some fidelity criterion for the process. This fidelity consideration is a good basis for distinguishing between the two types of source processing operations. Once this distinction is clear, the discussion can be confined to measures within the scope of the investigation being reported.

The source processing operations may now be distinguished by simply partitioning all possible compression transformations into two categories: those that are non-singular or strictly reversible, and those that are singular or irreversible. (A strictly reversible transformation of a signal is one that can be completely undone by a second transformation so as to yield a net result identical to the original signal.)

In a strict sense, the speech source is represented by a continuous analog signal, that cannot be described with finite sample data information. For transmission through discrete channels, the messages must be reduced to a discrete representation of finite length. Any operation of this type is irreversible in the strict sense. However, the measure, H, of source information rate is defined in a manner to recognize this situation. It is defined so as to explicitly depend on a fidelity measure. Henceforth in this report, source messages and channel messages are equivalent from the information theory point of view. The "information" of the finite discrete source representation, subject to the fidelity criterion, is the information to be conveyed by the channel message. The intelligence of source and channel messages are the same. The representation of the messages may differ from point to point, or with particular design, in a processing system, but the information theoretic message is the same. The irreversible operation of source processing is called signal conditioning, and is performed subject to some "tolerable" fiedlity degradation of the continuous source.

The source processing of messages are described so that all source coding operations are the strictly reversible operations in the information theoretic sense. Of course, in the broader than information theory view of the communication system, it is clear that absolutely perfect replication of the input at the output is not necessary (perhaps not even desirable) and that deliberately irreversible operations are a vital part of the overall compression process. Thus, consider a dicotomy of compression transformations; lump togethe~ all irreversible transformations under the title "Signal Conditioning", and all reversible transformations under the connotation of "Source Coding".

Without dwelling extensively on the design of signal conditioning transformations, one can at least pause to reflect on their relation to the overall source processing picture. As remarked earlier, there may be a number of different reasons why the total message bulk need not be communicated through a channel. These reasons are associated with different types of excess bulk, which in turn can serve to identify the type of transformation needed to eliminate this excess. Typically, one can describe the general objective of signal conditioning as eliminating message content which has no (little) utility. To illustrate, on a subjective basis.a portion of a message might be imperceptible, irrelevant, or unnecessary, and hence without utility. In speech, absolute acoustic intensity is irrelevant since the listener will probably adjust some volume control to this own liking anyway. The phase coherent waveform out of a vocoder analyzer filter is unnecessary, since a reasonable substitute can be provided by the synthesizer. And finally, any sound whose insertion or omission the listener could not detect is imperceptible and therefore without utility. If this type of excess content is to be eliminated from the transmitted message, it should be accomplished through some irreversible operation.

In contrast, source coding (reversible transformations) as here defined, has as its objective the elimination of only one type of excess message bulk. This excess is that part of any message that can be reconstructed by a knowledge of: the remaining bulk; the type of transformation that removes the excess; and certain characteristics of the process or mechanisms that produce the message. Note that the elimination of this type of excess, or so called "redundancy", must be accomplished with transformations in the representation of the message to reduce the apparent source bulk rate, or entropy. Fidelity, in the information theory sense, is to be retained in measure, not just on a subjective criterion of perception by the ultimate message sink.

The total of source processing as referenced earlier, has been defined as the composite operation of signal conditioning (pooling of bulk) and source coding (pooling of information). We are interested in the latter. An example for comparison accentuates this partioning of interest.

The difference in the output source bit rate of a vocoder as compared to standard pulse code modulation (PCM) coding of the input speech signal is a good example of reduction of source bulk rate through signal conditioning. In fact, it appears that the vast majority of prior speech compression efforts have been devoted to signal conditioning operations. The very phrase brings to mind a whole host of techniques which clip, chop, slice, and in other ways grind up speech waveforms to eliminate the parts that have no utility. Of course, the compression achieved has been less than ideal, since some useful message content is unavoidedly destroyed by most or all techniques. In fact, it has appeared at times that the process was "bottoming out" and that no further gain could be made without intolerable losses in fidelity. Nevertheless, throughout this intensive search the possibility that additional compression might be obtained by completely reversible, fidelity-preserving transformations seems to have been largely ignored except at a remote distance in theory. Although this motivation is by no means unique to speech compression work, application of source coding concepts to speech should be particularly productive. To restate it then, the basic purpose of this project was to investigate the application of source coding concepts and techniques of speech compression.

B. TERMINOLOGY AND MEASURES

For vocoded speech sources, and many others, the message bulk is available for transmission in the form of a sequence of symbols from a given alphabet. Compression of message bulk by source coding is the operation of transforming such sequences of symbols into a new second sequence of symbols, where the second sequence is smaller in some sense than the first. These operations are always defined here so that they have an inverse. In coding theory, the first concept to be grasped is that messages (entire sequences of symbols) are the basic entities produced by sources and operated upon by source coders, even though the coding operation is performed in time upon groupings of symbols within the entire sequence.

As an illustration, a speech sample of given duration is represented at the output of a vocoder analyzer by a finite length sequence of symbols in ordered groupings called frames. Consider the portion of each frame representing energy content at the sample time in the spectrum channels. In this instance, the symbols of interest represent amplitude levels of the spectral energy.

In all instances, the set of permissible values that a symbol can have is called its alphabet. Almost always, for considerations here, all symbols in a sequence will have the same alphabet. For the most part, we are interested in messages formed of symbols from a discrete alphabet. Development on this basis in no way constrains the generality of results, as is amply shown in the primary references.

The measure of message bulk (or bulk rate) most suitable for present purposes presupposes further notions about messages and sources that produce messages. In particular, the very nature of what constitutes information, and ones ability to measure information rate, must be described in terms of uncertainty. The uncertainty as to what message (or information) is to be communicated by a collection of symbols depends upon the probabilities associated with the selection or generation of messages at the source. It is this uncertainty that must be resolved by the correct reception of the message. Before reception, there is at most a probability associated with any one of a set of possible meanings implied by the symbol collection. Thus, the information content of the symbol sequence is to be defined in terms of the probabilities for all allowable messages. The source is described by (1) what messages it can produce, and (2) the probabilities of occurrence for each.

For each allowable message, X_α, that a source might produce, a probability, or probability density, P_α, is defined for its occurrence. Since any message produced must be one of the allowable set, the sum of probabilities, $\sum_\alpha P_\alpha$, is unity. The set, $\{X_\alpha\}$, and the associated probability measure, $\{P_\alpha\}$, is called an ensemble. For convenience, a finite message of length N + 1 symbols will usually be represented by the explicit symbol sequence, $X_\alpha = (x^\alpha{}_i, x^\alpha{}_{i-1}, \ldots, x^\alpha{}_{i-N})$. Similarly, the message probability is represented by the joint probability of the finite symbol sequence, $P_\alpha = P(x^\alpha{}_i, x^\alpha{}_{i-1}, \ldots, x^\alpha{}_{i-N})$. Also, the conditional probability, $P(x^\alpha{}_i / x^\alpha{}_{i-1}, \ldots x^\alpha{}_{i-N})$ is the probability that the symbol, $x^\alpha{}_i$, follows the N preceeding values, $\{x^\alpha{}_{i-1}, \ldots, x^\alpha{}_{i-N}\}$, in the sequence. Source ensembles for which communication problems (including the type we consider) can be described analytically are restricted to have certain statistical properties. Those properties will be noted as they are required.

In the preceeding paragraphs, signifiance of the representation of a message was discussed. It was noted that the information conveyed by the message was independent of the representation. In the following discussion source messages are assumed to be of finite length. The theory does not require this assumption, it is a matter of convenience for those not familiar with the subject. Even though the alphabet of message inputs to the standard discrete channel could be arbitrary, it is practical (to say the least) to assume a binary representation. The object of efficient source coding is to minimize the expected bit rate of this channel. The fundamental theorem of source coding[1] states that finite length messages, X_α, may be coded into uniquely decomposable binary sequences of length, ℓ:

where

(1) $\ell \geq H' = -\Sigma_\alpha P(X_\alpha) \log_2 P(X_\alpha)$. (1)

(2) ℓ may be made to approach H' arbitrarily close by coding message symbols in sufficiently long groups.

Several observations of this theorem are in order.

Uniquely decomposable means that output messages may be joined end-to-end without special punctuation and it is still possible to determine where one message stops and the next begins.

The lower bound on ℓ, H', depends only on the probability of occurrence of source messages and not at all on the alphabet of original representation. This measure, the now familiar entropy statistic, is the practical reference required to discuss minimum bit rates of a discrete channel. Information is thus measured by the minimum number of bits required to convey the uncertainty of the source message through the standard binary channel. The theorem as used here is not in its most general form, but adequately serves the purpose intended. Further comment on entropy as a practical measure will clarify the problem of ideal source coding.

In general, it is more convenient to use a measure of information per symbol, H, as a reference so that the length of messages is not explicity involved. For messages arbitrarily long, the average information per symbol in a sequence of length N approaches the average information per symbol. generated by the source as N increases indefinitely to the length of messages produced by the source. That is,

$$H = \text{Lim} - \frac{1}{N} \Sigma_\alpha P(x^\alpha{}_i, \ldots, x^\alpha{}_{i-N}) \log_2 P(x^\alpha{}_i, \ldots, x^\alpha{}_{i-N}) . \qquad (2)$$

This expression also requires comment.

The definition, Equation (2), is an expectation formed by averaging over all messages of the source ensemble. As is the case for most ensembles when using the entropy measure, averages just cannot be formed in this manner for speech sources. Thus, a hypothesis of the ergodic theorem[2] of probability (measure) theory is assumed so that an average over the ensemble of messages may be replaced by a time average over any one message of sufficiently long duration. This assumption must be remembered, since spoken messages certainly have structure that depends upon the speaker, or classification of speakers.

In the development of approximations for source entropy, the limit expressed in Equation (2) is truncated. In such instances the joint probabilities for the N symbol sequences depend explicitly on the time origin index, i. These same probability measures are used to derive source coding transformations. Again for practical considerations (at least in a preliminary investigation), it is necessary to assume that these probabilities are independent of the time origin of observation. A time series for which this assumption holds is said to be stationary.

These assumptions are restrictions on the source ensembles for which source coding transformations may be analyzed precisely. They are theoretical considerations, and as such are observed in practice only to the extent that one can assess (and is willing to lable) cause and effect. This assement is a typical, and difficult, association of model experience. One is tempted to say, "let the results speak for themselves!"

Equation (2), with its practical limitations, is the reference that serves the present need. The minimum possible channel bit rate implied by this measure can only be achieved (arbitrarily close) by rather complicated coding schemes. The fundamental channel coding theorem is an existence theorem that in one form or another appeals to the coding of long sequences of symbols. That is, a coding procedure of approximating arbitrarily close the definition of the source entropy. The description of such procedures is called ideal coding. It is important to establish bounds on the measure of source information rate, and a procedure for approaching an ideal coding to achieve any reasonable bound.

An attempt to measure the true entropy of an ergodic source would entail the gathering of statistics implied by Equation (2). This task would be of exceedingly great proportions for a source of any significant complexity. Rather than attempt the task in that manner, upper bounds may be established with probabilities more subject to practical measurements. These bounds are defined in terms of the conditional probability measure mentioned above, and the related marginal probabilities,

$$P(x^{\alpha}_{i-1}\ x^{\alpha}_{i-2}, \ldots, x^{\alpha}_{i-N})$$

defined by,

$$P(x^{\alpha}_{i}/x^{\alpha}_{i-1}, \ldots, x^{\alpha}_{i-N}) = \frac{P(x^{\alpha}_{i}\ x^{\alpha}_{i-1}, \ldots, x^{\alpha}_{i-N})}{P(x^{\alpha}_{i-1}, \ldots, x^{\alpha}_{i-N})}. \tag{3}$$

Any entropy measure determined by substituting one of these probabilities into an expression similar to Equation (2), without the limit of long sequences, will be called an "apparent entropy" (or information rate) to indicate that it is the entropy revealed or "made apparent" by the probabilities actually measured. This rate is defined subject to the ability to measure the relevant probabilities. That is, the uncertainty in the minimum channel bit rate is contingent upon what probabilities are actually known, or accurately estimated. In attempting to reduce this uncertainty to a minimum, it is convenient to establish criteria or references for use as comparisons.

The relative entropy of a source is conventionally taken to be the ratio of the true source entropy, H_{true}, to the maximum value, H_{max}, it could have while restricted to the same symbol alphabet. Maximum uncertainty is expressed by all symbols being equally probable, thus $H_{max} = \log_2 M$, for a discrete alphabet of M symbols. A convenient performance rating for any coding scheme is its compression factor defined by the ratio H_{app}/H_{max}. Thus, the relative entropy is the maximum compression possible by encoding into the same alphabet. One minus the relative entropy is the "redundancy" of coding with the H_{max} scheme. Any ideal coding is an attempt to approach the maximum message compression by coding to reveal H_{true}. In a practical situation H_{true} will seldom be known. Therefore, the goal of practical source coding is to make the ratio H_{app}/H_{max} as small as possible. Only estimates of how well H_{app} approaches H_{true} may guide these efforts.

By accumulating the necessary statistics for sequences produced by a source, one could attempt to measure a close bound on the true source entropy. If the limit in Equation (2) converges rapidly, the message representation is suitable for an efficient coding scheme. If long sequence probabilities are required for the apparent entropy to converge, no feasible, much less efficient, coding scheme of the "ideal" type is possible. The convergence of apparent entropy measures with increased sequence length depends upon the intersymbol influence of the message. Stated in more detail, speech source structure is known to contain elaborate organization over both time and frequency intervals.[5] This organization of the source is reflected in the signal (message representation) by constraints between symbols of the message. If symbols of the message were not constrained, that is, were statistically independent, the entropy could be measured and utilized by observing single symbol statistics only. The minimum possible bit rate could be achieved by coding symbols one at a time.

However, speech sources do not have this simple structure. The information produced by the speech sounds is spread over the message length. Any attempt to efficiently code the signal produced by such a highly organized, or "predictable," source requires that a large number of terms in the message be retained in a memory and coded at one time as a group. Probabilities must be accumulated for the occurrence of each such group. In addition, and more significant, a code book or equipment for identifying the assigning variable length codes must be part of the encoding system. For coding of sequences of N symbols from an alphabet of L levels, L^N entries are required for the code book. Substantial intersymbol influence in vocoded speech will be shown to exist over multiplexed sequences at least 36 or 54 symbols in length. For eight level quantization of channel amplitude values, code books with about 10^{32} or 10^{48} entries would be necessary — obviously an impractical state of affairs. Instead of multiplexing the channel samples, sequences from each vocoder channel could be processed by a separate encoder. For coding of channel sequences only two or three symbols in length, 18 encoders with 64 or 512 entries respectively would be required. It is a matter of record, that rather than go to this complexity and expense, inefficient single symbol statistics have been used with the simpler variable length encoding schemes. Or worse, maximally inefficient, and less complex, fixed length codes have been assigned on a uniform probability distribution over the symbol alphabet.

Certainly, for a great many communication systems similar to those of concern here, maximum efficiency of transmission is just not a great concern. For high priority digital links where transmission efficiency is of concern, cost and complexity may be justified (within reasonable bounds) for sophisticated processing. More generally, however, it is argued that channel bandwidth or power may be increased more easily and possibly with less expense; alternatively, the acceptable limits on fidelity are reduced so that inefficient coding techniques remain in vogue. These arguments have merit in many instances as far as the numbers trade-off is concerned. On the other hand, considerations regarding maximally efficient utilization of existing equipments may reach different conclusions. In any event, for the present or for the future, a fresh approach to the ideal coding has been developed, and is described in the framework of the terminology presented herein.

C. AN APPROACH TO IDEAL CODING

The primary concept of approaching ideal coding by predictive transformations of the message representation may be stated quite briefly. Rather than observe and utilize probabilities for long sequences, an attempt is made to transform the message presentation in a manner that includes as much as possible the influence of these probabilities, resulting in a new representation with lower apparent entropy in short sequences. In particular, the apparent entropy of major concern is that of the single-symbol probabilities. It is desired that as much as possible of the source redundancy be transformed to the single-symbol statistics, not over sequences of greater length. These concepts and procedures are iterated and reiterated with the hope that some intuition and insight may result.

The true entropy of a source is bounded by two forms of measuring apparent entropy. The first is represented in terms of the entropy of marginal distributions of an N + 1 symbol sequence. The inequalities,

$$H(x_i \cdot x_{i-1}, \ldots x_{i-N}) \leq H(x_i) + H(x_{i-1}, \ldots, x_{i-N}) \leq \ldots \leq H(x_i) + H(x_{i-1}) + \ldots + H(x_{i-N}), \quad (4)$$

are an expression similar to what has been said previously. The equalities hold only if the message symbols are statistically independent. At any stage in extending the probability measurements over longer sequences, the entropy can be no worse than that computed with shorter joint probability observations. The second inequality,

$$H(x_i, x_{i-1}, \ldots, x_{i-N}) \leq H(x_{i-1}, \ldots, x_{i-N}) + H(x_i / x_{i-1}, \ldots, x_{i-N}), \quad (5)$$

has a similar interpretation, but is defined in terms of the Nth order conditional distribution. Combining the two forms of inequality concerning the entropy of the sequence, there results,

$$H(x_i / x_{i-1}, \ldots x_{i-N}) \leq H(x_i), \tag{6}$$

which shows that the entropy of the symbol, x_i, based upon a knowledge of past symbols in the sequence is a better (at least can be no poorer) measure of the average information per symbol than the entropy based upon the first-order marginal or single-symbol statistics. For organized sources, the strength of the inequality increases with increased N, asymptotically to the limit over which intersymbol influence exists. The implication of this relation is that rather than attempt to code long sequences to achieve ideal coding, one could attempt to predict a present symbol by utilizing an approximation to the conditional distribution dependent on past symbol values in the sequence. This process could be disasterous for some sources. A poor approximation could give a transformed symbol value containing all the information of poor prediction. However, for sources with message statistics such that a reasonably simple and accurate approximation function exists, there is the potential of a good bound on the true entropy. In fact, one can improve the estimate of this bound by a further consideration

The nature of the measure of information (rate) leads directly to the Averaging Theorem, stated simply as follows. Any transformation of the symbols (or sequence of symbols) that tends to equalize the probabilities of occurrence of the symbols (or sequences) will produce an increase in the entropy of the symbol(or sequence). The converse holds, and is of primary concern here. If the probabilities of occurrence can be "peaked" within a symbol alphabet, the bits per symbol required to transmit the information of a symbol or sequence can be decreased.

Conceptually then, in designing practical source coders, it is desirable to perform transformations that peak the probabilities over the new symbols and thereby decrease the apparent entropy. Transformations of this sort must preserve the true source entropy. Recall that the apparent entropy of a symbol (or sequence) is that which can be measured as a bound on the true entropy. The lower the apparent entropy, the better the approximation to the relative source entropy, and the more the compression available in the redundancy of the source is achieved. Also, to the extent that intersymbol influence is removed, the transformed symbols are independent, and accordingly, are easy to code efficiently.

To summarize at this point, recognize that: (1) ideal coding to approximate true source entropy of an "organized" source entails the consideration of coding with long sequences over which the intersymbol influence exists; (2) the large code-book and statistics of distributions over long sequences prohibit the direct approach of the definition of information for coding of most organized sources (that is, sources of interest imply highly redundant first-order coding);

(3) the redundent information of organized source messages that is evidenced by the intersymbol influence of long sequences may be removed by sufficiently accurate transformation that utilize the constraints of the conditional probability distribution of the original sequences; (4) transformation of this sort can be realized, at least conceptually, by implementing techniques that preserve the true source entropy but "peak" the symbol probabilities thereby decreasing the apparent source entropy; and (5) as the apparent source entropy of the transformed message decreases to a close bound on the true source entropy, the redundancy of this representation approaches the minimum, implying that excess bulk has been removed in the information theory sense There remains the choice of a form of transformation that can be implemented and will accomplish the results just described.

D. PREDICTIVE CODING TRANSFORMATIONS

The previous sections were intended to motivate what at the beginning was labeled source coding. Source coders are transformations of the message representation for purposes of removing "redundant" message bulk. That is, a compression of bulk is to be achieved by representing the message in a maximum information per symbol manner. The design of source coders is a two-step process. First one must discover and measure the statistical constraints which govern the production of symbol sequences by sources, and then develop means of removing the predictable (redundant) message content.

From one point of view, identifying the predictable structure of the source ensemble is the easy part of the procedure because each and every characteristic of speech that makes it look or sound different from band-limited white or flat noise constitutes a constraint on the source distribution and thus, indicates a potential compression. However, from the very practical point of view of the problems discussed with regard to ideal coding, the statistics necessary to identify all redundancy (measure true entropy) are not simple to accumulate. Thus, a restriction is imposed on the "kinds" of predictable structure that can be recognized. It follows, that the transformations for removing redundancy are likewise limited.

No practical theory of source coding has previously been developed. The approach that has been proposed here for reducing the apparent entropy of a restricted message ensemble is in essence a transformation of message representation that utilizes a knowledge of past symbol values in a sequence to predict a current value, or more specifically, the "expected" current value. A peaking of probabilities over the symbol alphabet is to be realized by coding only the error between the actual and the predicted symbol value.

There is theoretically no limit on the form of transformation (classes) that could be used for predictive purposes. The choice of any one form is dictated by the feasibility of obtaining statistical information necessary to evaluate its merits, and the complexity one is willing to accept in analysis. There are limited classes on nonlinear forms for which analysis of choosing the optimum member may be performed with reasonable effort. These classes are primarily nonlinear means of combining linear analysis, and are not worth general discussion. The simpler classes of linear transformations have been used to pursue the analysis of this investigation. For this limited class, techniques of analysis are available and the necessary statistical measurements turn out to be most reasonably acquired. As one might observe in Section III, even the simple linear analysis can be difficult due to sheer magnitude of size.

With a limitation on the class of acceptable transformations, it is rather obvious that the optimum function in the class is the one that results in a minimal apparent entropy representation of the transformed message ensemble. Here also, there is more to be considered. The criterion of minimum H_{app} may be difficult to apply directly toward choosing the optimum transformation of a given linear form. In fact, a very interesting, and messy, variational problem exists for describing this situation. As a simpler (and in most cases just as effective) criterion, the form of linear transformation is chosen so that the errors in prediction are a minimum in the root mean square sense. This criterion provides a "best" predictor in the sense that the expected prediction error is zero, intersymbol correlation is removed, and the degree of peaking of the distribution about a single symbol probability is explicitly maximized. The latter follows from the minimum variance (minimum rms error) property, which also implies a minimal entropy (uncertainty) representation in the absence of other information concerning the source ensemble.

The transformation scheme may be represented in the form,

$$y_i = x_i - p_i, \tag{7}$$

where y_i denotes a symbol of the transformed message representation, x_i represents a current symbol value, and p_i is a number computed from the linear form,

$$p_i = \sum_j c_{ij} x_{i-j}, \tag{8}$$

as a prediction of the value of x_i based upon past symbol values. Note that the objectives of both transformations described qualitatively in the motivating remarks are incorporated in this single transformation. For a best rms sense prediction of the mean of the conditional probability distribution,

$$p_i = \overline{x}_i = E(x_i/x_{i-1}, \ldots, x_{i-N}), \tag{9}$$

it is easily shown that the resulting error term distribution has minimal variance and zero mean. Since a transformation of this type represents a simple translation (no change of scale) in the mathematical sense, true entropy is preserved over the discrete alphabet. Also, in the terminology of statistical regression analysis, a predictor of this type is "best" in the maximum likelihood sense if the resultant errors are normally and independently distributed. It will be seen later that these conditions are for the most part satisfied with prediction models utilized in the study Verification of the independence condition would imply that the transformation yields a new sequence with apparent entropy closely bounding the true source entropy. There are several other desirable consequences of this predictive transformation that occur as by-products of the design criteria A discussion of these features is incorporated in the analysis of results.

Properties attributed to the predictive transformation could just as well be discussed in the terminology of Markov processes. An N^{th} order Markov process is a stochastic process for which each state (or symbol, in our application) depends only on the N previous states (N past symbol values) of the process. This terminology is mentioned in passing, since a process of this type is easy to visualize, and any approximation of the message conditional probability distributions is an approximation with confidence measures expressible in terms of Markov processes. No great benefit is derived by phrasing the analysis in this terminology, but a reader familiar with such processes is alerted to the parallels.

The primary objectives of this investigation were pursued subject to the theoretical motivation and limitations expressed in this section. The remaining sections of this report must necessarily describe how physical reality evolved from conceptual model.

SECTION II

FORMULATION OF PREDICTIVE CODING SYSTEM

A. LINEAR LEAST SQUARE PREDICTION

Before considering in detail the block diagram of a (transmitter-receiver) predictive coding system, a detailed familiarity with the component blocks is helpful. The predictor function is of most immediate interest. Along with the motivation toward simplicity of analysis and hardware implementation, the linear approach to prediction is a logical "first-step." A first approach to nonlinear functions may be considered a correction process applied to a linear approximation. With or without this generalization in mind, the effectiveness of the linear attempt needs to be well understood first. If the simpler techniques can accomplish the job (and can be evaluated), it is unnecessary to proceed with further complication.

The theory of time invariant linear least square (LLS) prediction dates back to the early work of Gauss, and the least square norm is fundamental to most of the mathematical theory of approximation. In applied statistics, the LLS methods of regression analysis constitute the backbone of interpolation and extrapolation analysis, with areas of application far too numerous to mention. More recently, the Wiener[2]-Kolmogoroff[4] formulations of prediction and linear filtering* of stationary time series have combined and augmented results from statistics and functional analysis, and have firmly implanted this theory in the signal processing analysis of communications. The ground work is more than adequate, the practical problems of effecting the analysis remain.

It will be shown in Section III that the source statistics necessary to compute linear predictors are estimates of the discrete set of time series autocorrelation coefficients.

$$\phi_k = \lim_{L \to \infty} \frac{1}{2L + 1} \sum_{i=-L}^{L} x_i x_{i-k}. \tag{10}$$

The estimates are sample measures of the variance and co-variance between message symbols separated by k terms in the time series. The sufficiency of these statistics for LLS prediction using N past symbols is the overwhelming simplicity that results compared with measures necessary for approximating an arbitrary nonlinear function of N variables.

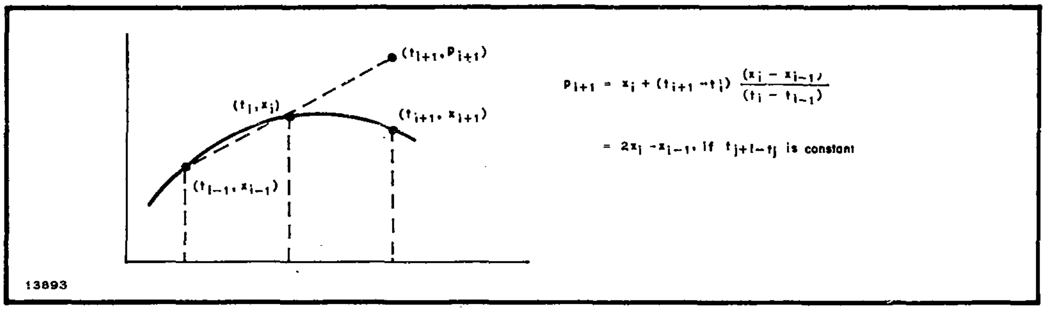
*A major proportion of this work is equivalent to efficient filter design for vocoders. An experimental implementation in hardware can be expected to develop along those lines.

Both the theory and measure of source statistics for LLS prediction require an assumption of stationarity of the time series. That is, the function that relates the immediate past N symbols in a sequence to a new value must be independent of the position of the N + 1 symbol group in the overall message. The degree to which this assumption is satisfied can influence the prediction effectiveness far more than the constraint of linearity. In particular, it will be noted that a single LLS predictor for a multiplexing of all vocoder channels (treated as a single time series) is not effective, but with separate predictors for each channel, the stationarity assumption is reasonably satisfied.

As a final comment at this time concerning the source statistics, one may think of speech source ensembles in many ways. The ensemble consisting of all messages from a single speaker is the most reasonable, but limited, class to discuss. A second class might be a single message spoken by any male or female speaker. Any arbitrary message spoken by any male or female speaker would be an enlargement of that class. Or, the ensemble of all messages spoken by all male or female speakers would include all three previous classes. Obviously, for communication purposes, the more inclusive the ensemble over which a given statistical predictor is accurate, the more economically sound would be the expense of sophistication in signal processing. At the other extreme, the more restrictive a measure is over a set of classifications, the more effective it is for identification purposes. Although it is to be desired, from the communication viewpoint, that statistical prediction would be effective over a wide source classification, there is merit in a good prediction function over limited ensembles, if the classification categories can be identified and the prediction function easily adapted from one category to another. This connection between the predictor form and the source statistics must, and will, be accorded further comment in the evaluation.

B. NONSTATISTICAL PREDICTION

A second form of prediction, totally unrelated to the major motivating arguments, should be mentioned for completeness. Prediction, of a sort, is the objective of a multitude of nonstatistical interpolation and extrapolation techniques fundamental to mathematical analysis. For example, consider the linear single point extrapolation of values in a discrete data series. By assuming a constant extension of the straight line slope between the immediate past two known values, the approximate next value may be computed as indicated in Figure 2. The model is crude, certainly, but it requires very little computation. It is also a matter of record that in similar applications, this simple technique has frequently been as effective as the more complicated approach described above. Slightly more advanced forms are based upon assumptions of constant curvature, higher polynomial approximation, gaussian weighted estimates, etc. — a vast and useful theory for appropriate applications.



$$p_{i+1} = x_i + (t_{i+1} - t_i) \frac{(x_i - x_{i-1})}{(t_i - t_{i-1})}$$

$$= 2x_i - x_{i-1}, \text{ if } t_{j+1} - t_j \text{ is constant}$$

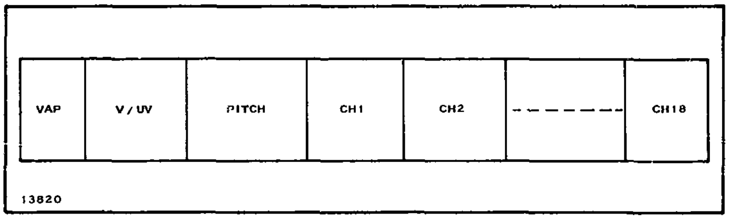
Figure 2. Linear Extrapolation

The present interest, for the most part, does not qualify for such an application. Preliminary investigation showed that most of these techniques result in the aforementioned disaster area — the error sequence containing all the increased apparent entropy of poor prediction.

As a possible exception to the conclusion of the last statement, one nonstatistical predictor seemed worthy of further consideration; this resulting when one accepts the immediate past value of a symbol as the "best" estimate of the current value. Again, the computation is utter simplicity. This simple "differencing" scheme can be followed by a transformation of the LLS statistical type. The analysis and evaluation of this prediction scheme was included in the investigative program, and will be discussed in detail.

C. DESCRIPTION OF VOCODED SPEECH SIGNALS

The analog speech signal produced by a high quality dynamic microphone can be sampled and processed by a vocoder* to produce a frame sequence with the character format illustrated in Figure 3.



VAP	V/UV	PITCH	CH1	CH2	-------	CH18

Figure 3. Vocoder Frame Format

*Speech data from the 18 channel research vocoder at the Data Sciences Laboratory, Air Force Cambridge Research Laboratories, was used for this investigation. The format described here is for that vocoder, but representative of other designs except for the spectrum normalization factor, VAP.

The voice frequency range, 70-4000 cycles, is typically analyzed and synthesized ıor representation of energy in 18 spectrum channels, such as indicated in Table I. The channel data used in this study consisted of an eight-level, logarithmic, quantization of spectrum amplitude values, following a normalization process unique with the AFCRL vocoder. A 50 frame per second sampling rate was used for all data. The VAP format character is the spectrum normalization factor, and is discussed later in this section. The V/UV character labels the frame spectrum values as representating voiced (V) or unvoiced (UV) speech. The PITCH characters represent the fundamental pitch frequency. The VAP and V/UV characters may be utilized as a basis for prediction using different source statistics, that is, predictors computed from voiced only, unvoiced only, or nonsilence only speech statistics.

In the investigation of predictive coding, only the processing of channel spectrum data has been analyzed. It is assumed that similar coding of the remaining frame data would result in further compression of the total vocoder bit rate. However, this data will vary more with equipment design, and change with the improvement of new designs, and may be processed separately if multipurpose use is made of transmission channels, for example, in the multiplexing of teletype data.

D. PREDICTIVE CODING COMMUNICATION SYSTEM

The operations, constraints, and performance criteria discussed to this point are now related to a typical processing-block diagram of a system for communicating vocoded speech (see Figure 4). The diagram should be self explanatory, but the following brief remarks are possibly of interest.

A description of the processing at the transmitter describes the reverse operation at the receiver. The previously described frame data represents the output of the vocoder analyzer. In one form of analysis, the 18 channel amplitudes are differenced on a frame to frame basis before processing. As a convenience to analysis, and in no way reducing the accuracy or generality of results, the statistical mean amplitude values were subtracted from the channel outputs in the other processing models. This operation represents a translation of origin in the mathematical analysis and a voltage reference in any equipment. As indicated, in any analysis or processing scheme, the separate channel amplitude values in any frame are multiplexed into an arbitrary serial order. The predictive transformation discussed in Section I is accomplished by the subtraction of the predicted estimate of a symbol from its actual value. Since the form of magnitude representation (analog or digital) is rather arbitrary at this point, it is assumed that the analog or many-level quantized (quasi-analog) difference must be quantized for presentation to the Huffman Coder.

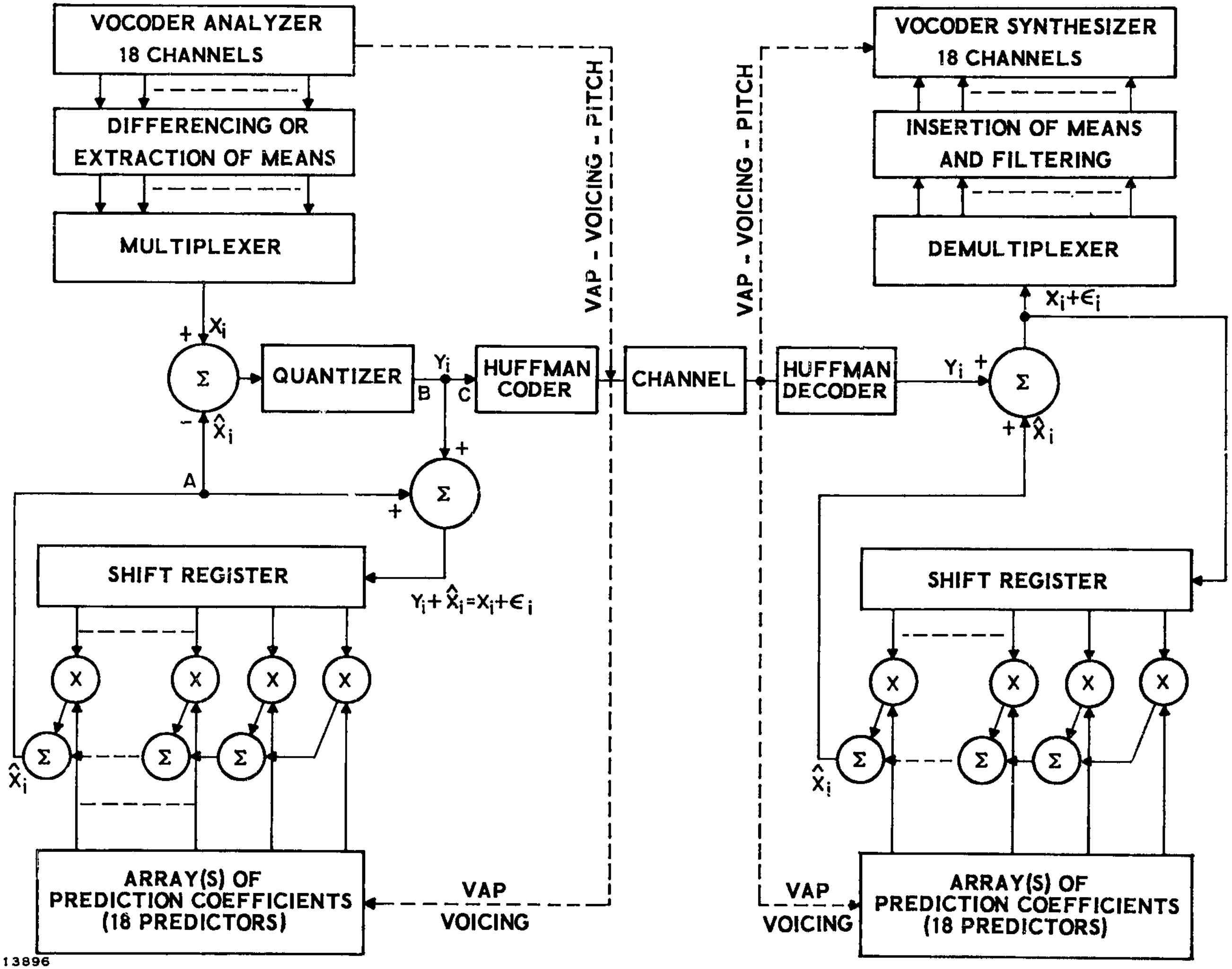


Figure 4. Predictive Encoding System-Block Diagram

Table I. Center Frequencies and Bandwidths for Conventional Mode of Operation

Filter No.	5-DB Points			Center Frequency (f_o)	3-DB Points		
	f_1	f_2	BW		f_a	f_b	BW
0	70	200	130	118.321	80.97	172.89	114.320
1	200	332	132	257.679	215.20	308.54	116.079
2	332	464	132	392.488	348.58	441.92	116.079
3	464	596	132	525.872	481.27	574.61	116.079
4	596	728	132	658.700	613.68	707.02	116.079
5	728	860	132	791.251	745.96	839.29	116.079
6	860	992	132	923.646	878.15	971.49	116.079
7	992	1132	140	1059.679	1011.34	1110.33	123.114
8	1132	1287	155	1206.985	1153.43	1263.03	136.305
9	1287	1460	173	1370.759	1310.96	1433.29	152.134
10	1460	1650	190	1552.084	1486.36	1620.71	167.083
11	1650	1859	209	1751.392	1679.06	1826.84	183.792
12	1859	2089	230	1970.624	1890.98	2053.62	202.259
13	2089	2342	253	2211.860	2124.22	2303.12	222.484
14	2342	2619	277	2476.632	2380.63	2576.50	243.590
15	2619	2921	302	2765.881	2661.17	2874.71	265.574
16	2921	3251	330	3081.618	2967.15	3200.50	290.197
17	3251	3610	359	3425.776	3301.20	3555.05	315.699
18	3610	4000	390	3799.987	3664.60	3940.37	342.960

Note that the Huffman Coder is but one of many coders that could be considered to fit the needs and the concept of "standard channel" introduced in Section I. It happens to be a highly efficient variable code length assignment scheme, with prefix and other properties to be described later. It also contains the active memory (buffer) and delay logic necessary for implementing an efficient variable-length code, and will serve as a model for investigating the extent of these requirements. Further comments on Huffman coding are made later in this report.

To satisfy the requirements for a totally reversible process in the absence of channel noise, the sum of the quantized difference and the original prediction are the actual values used in the "feedback" type process for prediction. As indicated on the diagram, and discussed extensively in this section, the ϵ_i represent the "quantization noise," or difference between the value of an original symbol and its approximate value actually utilized in the prediction of future symbol values. Observations concerning this ϵ_i noise serve for a comparison of the "fidelity" of the transformed message with the unprocessed original data input. Any possible psychoacoustic degradation of quality resulting from ϵ_i is indicated at the output of the inverse processing at the receiver.

The essence of processing for predictive coding is diagrammed without the features that distinguish one prediction scheme from another. To allow for this flexibility, VAP and voicing information are sprinkled liberally throughout the diagram, and additional such lines of logical interest may be added as the models are developed and discussed in Section IV.

E. FIDELITY CONSIDERATIONS REGARDING A PREDICTIVE CODING SYSTEM

In Section I, the overall source processing operations were rigidly, but arbitrarily, partitioned into two categories: the irreversible signal conditioning operations, which determine system fidelity, and the irreversible source coding operations, which exploit source redundancy but have no effect whatsoever on system fidelity. The objectives in doing this were: first, to emphasize the basically different considerations involved in the two types of compression operations, and second, to simplify the exposition of the principles of source coding. Implicity,. through the diagram in Figure 1, it has been suggested that the two types of operations could be implemented in two distinct subsystems connected in series. While a physical system could undoubtedly be organized in this manner, there is really no commanding reason for doing so. On the contrary, in a practical system utilizing predictive encoding, there are very definite advantages to arranging the component operation of the overall source processing system in a less disjointed fashion. To be specific, it is desirable to relocate the quantization process, usually the last step in signal conditioning, as an integral part of the prediction procedure. However, when this latter order is used, the net or effective signal conditioning procedure is unavoidably changed when compared to the former order. Consequently, fidelity considerations must be reintroduced into the discussion of source coding operations. Three questions need to be discussed in this section: (1) the source processing configuration most desirable from an overall system effectiveness point of view, (2) the manner in which this configuration alters fidelity as compared with a straight "serial" configuration, and (3) the study questions generated in connection with this modification relative to the idealized system that are suggested by the discussions of (1) and (2).

To anticipate a possible misunderstanding at this point, it should be noted that the encoding process per se is still required to be strictly reversible. The suggestion is that the signal conditioning operation can be revised slightly to improve the effectiveness of the combined compression operation. Although a true determination of fidelity can only be achieved through psychoacoustic methods, there is good reason to believe that the "modified" system would at least preserve, and would probably enhance, the fidelity of the present AFCRL vocoder.

In order to have the encoding process be reversible, the prediction, p_i, must be computed via a closed loop or feedback procedure. To realize this, one observes that the predictions developed at the receiving end of the system must necessarily be computed from the error signals, y_i, because this is the only data available. Clearly then, it is required that the predictions developed at the transmitter be exactly equal to those produced at the receiver. One way to ensure this is to simply duplicate the receiver's processing at the transmitter. This scheme is illustrated by the feedback procedure shown in Figure 4, which is used in both the transmitter and the receiver. Although it may not be obvious at this point, it may be shown that a feedback procedure of this general type must be used if both coder and decoder are to have finite memories.

This feature of reversible predictive encoders is necessary regardless of whether the signals involved are analog or digital. However, let it be supposed for now that the input in Figure 4 is a sequence of integers, and suppose also that the output is to be fed into a Huffman Coder designed to accept sequences of integers. Such a system is not reasonable, however, because the predictor coefficients cannot be derived optimally as integers. That is, for optimum predictor coefficients, neither the prediction, p_i,nor the error signal, y_i,will be integers, and the error signal is incompatible with the input requirement of the Huffman Coder. To correct this situation while still retaining the desired simple form of the predictor itself requires that other system components be modified.

One possible solution is to simply round off a prediction to some integral value at point A in Figure 4. Because the same round-off rules can be used in both transmitter and receiver, the resultant encoding process is reversible. This scheme has a "philosophical" advantage in that it precisely fulfills the idealized demand imposed on source coding in Section I. That is, it is a reversible process, having no effect whatsoever on system fidelity. It has the disadvantage however of lessening the effectiveness of the prediction. Further, it is contingent on the assumption of integral valued input.

Insofar as the compatibility of the error signals with the Huffman Coder is concerned, the round-off procedure could just as well be located at point B or C of Figure 4. In either of these positions, it is easy to see that there is a direct path along which the round-off noise can pass so as to appear in the final output and possibly degrade overall system fidelity. Setting this vital aspect aside temporarily, consider just the effects of choosing either location B or C.

First, it should be clear that the entropy of the error signal is directly increased by the addition of round-off noise at either point B or point C. Therefore, the average bit rate of the Huffman Coder output will be increased in either case. Next, in addition to this direct effect, note that a round-off operation applied at point C in the transmitter cannot degrade prediction accuracy as it can when applied at point B. That is, in case B round-off errors can circulate around the loop indefinitely, whereas in case A they never enter it. From this observation, one might hastily conclude that point C is the preferred point at which to perform round off. To show otherwise, the following consideration of how round-off error affects the prediction in the receiver loop rules more or less decisively in favor of rounding-off at point B.

Clearly, when round-off is performed at point C, the predictions produced in the encoder and decoder will differ. In the former device they are computed from the unrounded error signals, and in the latter they are computed from rounded error signals. It follows, that in this case round-off error can accumulate indefinitely in the receiver loop with the result that the receivers predictions amount to nothing more than random numbers, essentially independent of the source signal being encoded at the transmitter. The result is an irreversible process, and rounding-off at point C can have a completely devastating effect on system fidelity. Although there are techniques whereby the build up of round-off errors in the receiver loop can be controlled, these techniques usually require greatly increasing the size of the memory used in the receiver. To all intents and purposes then, one is obliged to insert the round-off procedure at point B. When this is done, it may be noted that the predictions produced at each end of the system are identical. Except for the round-off error which is added only to the final system output, the net encoding scheme is reversible.

Consider now what can be done about the round-off error and its effect on system fidelity. Recall that at the beginning of this section it was assumed that the input sequence to the source coder was a sequence of integers. Obviously, the entire question of round-off error stems directly from this assumption. As was noted, this sequence of integers was in reality produced from analog data by a quantization procedure performed as the final step in the signal conditioner. It should be clear that quantization and round-off are nearly identical operations, and should have nearly identical effect on system fidelity. In fact, so similar are the operations and their effects, that round-off is sometimes referred to by others as "the second quantization." When viewed in this fashion, it is natural to ask if the system could be designed so that just one "quantization" process would be required. In answer, it is immediately apparent that a quantizer preceding the source coder may be eliminated and a new "equivalent" one may be inserted at point B. This arrangement is the final step in arriving at the coder configuration assumed as a model in this study.

At this point is is tempting to simply identify the quantizer as the only source of error in Figure 4, and then to assert that the source coding has been added without affecting system fidelity. Of course, with the signal conditioning and source coding operations thoroughly scrambled together as they are in this model, it is rather meaningless to try to attribute resultant system errors produced to either process. More to the point, it must be admitted that the precise nature of the quantization performed on the signal has been changed by moving it to a new location. The psychoacoustic testing is the only true test of system fidelity. Except for a few comments, the reader will readily be able to interpret for himself the probable effects of the processing system configuration on fidelity.

Two effects must be considered. These effects arise because real quantizers, such as that in the AFCRL vocoder, actually perform two operations on the input signal. Besides "rounding-off" analog values to nearest of several discrete values, real quantizers also clip large excursions of the input signal.* To explain the changes made to these two operations when the location of the quantizer is changed, it is convenient to consider the operations separately. First, ignore the clipping effects by considering quantizers which have a number of levels so large that the input signal is never clipped. If one assumes that each quantization interval has the same width, and the value assigned to an input value falling within a given interval is the mid-point of that interval, then the concrete statement can be made that the magnitude of the quantization error never exceeds more than one half the width of the quantization interval. Figure 4 illustrates that the subtraction of the prediction, p_i, from an input value, x_i, is simply a scale preserving translation

*Even when systems employ an AGC capability, there remains some clipping of peak signals. This is particularly true for peaks occurring in shorter time periods than the time constant of AGC circuit.

of the origin from which the value x_i is measured. Consequently, if a quantizer (which does not clip) is moved from a position immediately preceding the source coder to point B within it, the maximum possible quantization error is not changed. Thus, the only way in which the relocated quantization process differs from the original is through the continuously varying position of the intervals with respect to some absolute reference point. In most utilizations of quantizers, fidelity depends only on the relative width of the quantization intervals and not on their absolute position. It seems highly improbable that a listener will perceive this effect.

The changes in the clipping operation itself, and in its effect on fidelity, are both more relevant and more difficult to assess in advance of experiment. In contrast with the above, the effect on fidelity resulting from a clipping operation can be expected to depend significantly not only on the number of quantization intervals between the clipping points, but even more on the precise location of these points with respect to the absolute reference to which the input signal is measured. Thus, the prediction process, which effectively results in a time varying translation of the quantization intervals with respect to the absolute reference, will probably have a pronounced desirable effect on fidelity. A few observations may be made concerning this effect.

Certainly one can expect system fidelity to depend upon the frequency with which input signals are clipped. Insofar as this aspect of clipping is concerned, one can expect the prediction process to improve fidelity by reducing the number of times an input signal suffers clipping. To understand why this should be the case, recall the earlier discussion concerning the probability distribution of the original input symbol, x_i, and the error symbols, y_i. (In preceding paragraphs, these symbols have been considered to be discrete, or digital, numbers. The argument applies equally well to the present case where the symbols are continuous, or analog values.) The first order marginal distribution of the input symbols is relatively broad. The probability that an input symbol will be clipped is represented by the area under the "tail" of this distribution beyond the clipping point. In contrast, if the prediction process has been even moderately effective, the distribution of the error symbols will be relatively narrow and will have most of its area concentrated between the clipping points, usually around the midpoint. Thus, the probability that an input symbol will be clipped should be less when the transformed signal is quantized than the probability when the input symbols themselves are quantized. This effect has been demonstrated and measured in the experimental program.

In passing, it should be noted that the influence of clipping also depends significantly on the exact time at which the clipping occurs. Certain parts of the speech waveform are more important to the listener than others. In this regard, there could exist a tendency for the modified configuration to clip frequently at the transition points between phonemes, if prediction accuracy deteriorates significantly at these times. This question has not been studied experimentally.

Finally, there is one last aspect of clipping that requires comment. It is explicitly assumed that the width of the quantization intervals were kept the same when the quantizer was relocated. Implicitly at least, it is also assumed that the number of levels should remain the same. There is, however, no commanding reason for imposing this requirement. In retrospect, it seems clear that the choice of precisely eight quantization intervals (used in the AFCRL vocoder from which data was obtained) instead of say, 7 or 9 intervals, resulted from the desire or assumption that conversion to three binary digits would follow the quantization. However, where a variable length coding system is to be employed, any number of quantization levels may be used. In particular, the number of levels may be adjusted as necessary to obtain a desired degree of fidelity with little effect on the output bit rate if predictive coding is utilized. Further, with the unimodal distribution resulting from the predictive transformation, it should follow that the number and width of quantization levels may be adjusted in a very controlled manner. Because, as a result of the prediction process, the error symbol distribution should be highly peaked compared to the input symbol distribution, increasing the number of quantization levels should also result in a much smaller increase in the resultant average bit rate from the rearranged system than from the original. This effect has also been studied experimentally.

F. LIMITATIONS IMPOSED BY SAMPLE SOURCE DATA

As always seems to be the case, the speech samples used in this study impose certain limitations on the precision, significance, and extrapolation of the experimental results. For the most part, these limiting characteristics of the data are either discussed in othe[illegible]tions or are left entirely for the reader to infer. In this section, four characteristics not covered elsewhere are discussed.

First, there is the fact that the speech samples used were already clipped and quantized to eight levels, whereas the coding system being analyzed would preferably operate on the analog signal as it appeared before either of these operations. In the experimental work there was no choice except to treat the quantized values as if they were the actual sample values of the original analog waveform. Insofar as the steps leading to the design of the predictors themselves are concerned, the fact that the raw data has been quantized probably has little or no effect. The clipping on the other hand, certainly does have an effect on this portion of the results, but without knowledge of the appearance of the speech waveforms prior to the clipping, it is difficult to comment, even qualitatively, as to its nature. In the second portion of the experimental work, the evaluation of the net encoding process through direct simulation revealed that quantization of the original data does have an effect. As was noted in the preceeding section, quantization introduces a type of random noise into the sample data. This noise inevitably increases the entropy of the signal being coded so that the average bit rates actually observed in the simulation are somewhat larger than those which would have been achieved with the original waveform. How clipping of the raw data affects the simulation result is indeterminate. These effects will be considered again in the evaluation section.

The second characteristic of the speech samples to be considered here is the spectrum normalization feature of the AFCRL vocoder. "Spectrum normalization"[6] transforms the spectrum pattern to a form that is independent of the gross voice amplitude. This process is reversible when accomplished in the following manner. The analog output from each of the channel low-pass filters is summed at each frame to produce a signal proportional to the total amplitude of the input voice signal. The sum is called the voice amplitude parameter, VAP. It is then used as the scale-determining (variable) reference voltage in an analog to digital converter. In effect, each resultant spectrum analyzer channel output is inversely proportional to VAP. In this normalization scheme, a VAP value of zero for a speech data frame implies that all channel amplitudes in that frame are zero. This result is of considerable importance in the design of some of the prediction schemes and processing models.

The value of such normalization has been adequately demonstrated, and it seems certain that the technique will be incorporated in future vocoders. Many existing systems, however, do not have such a feature. The objective at this point is simply to warn the reader against incautious application of the estimated bit rate compression factors to these systems. It is exceedingly difficult, if not impossible, to establish analytically whether source coding would be more effective or less effective on unnormalized signals. At this time, it appears that good heuristic arguments can be constructed to support either conclusion, the weight of the moment favoring more effective results, since the amplitude constraints of the source (removed by normalization) are conceivably predictable by the source coding model.

The manner in which the speech samples were originally generated is the third aspect of the speech data sample that inevitably limits the utility of quantitative results. Specifically, the original voice recordings were prepared by having talkers read aloud from lengthy prepared printed text. Two excerpts from these recordings were selected for use in this study. The same basic text was read by each talker, but that is not the sole condition to be of concern. (The experimental results tend to substantiate the original assumption that choice of specific texts material should not be significant if sufficiently long samples are used.) It is the simple fact that a prepared text was used that is of importance. As a matter of contrast, in an actual communication application the coding procedure would be operating on so-called "conversational" speech. It follows from the following consideration that this is a difference of major importance.

When an individual talks there are short pauses or momentary intervals of silence interspersed throughout the speech waveform. Many of these, such as the pauses between sentences, are clearly audible even to an untrained listener. In addition to these obvious pauses, the so-called "connected" or continuous part of the overall speech waveform also contains numerous shorter, and usually inaudible, silent intervals. All in all, a very substantial percentage of an entire speech signal is dead silence! The present concern originates with the fact that the exact value of this fraction varies drastically with the conditions under which the talker is operating. The fraction may vary from something less than 25 percent for a highly agitated or pressured talker, to something well in excess of 75 percent for a talker about to fall asleep at the mike. Generally, one can assume (and has been observed), that when reading from a text, talkers produce substantially less silence than do talkers engaged in "normal conversation."

The compression factor that can or will be achieved by predictive encoding depends heavily on the amount of silence on the input waveform. In fact, of all the parameters that determine the average bit rate resulting from coding, the silence-nonsilence ratio is by far the most significant. One is thus faced with the rather ticklish problem of converting experimentally observed coder performance based on one type of speech data into predictions of the performance to be expected in an as yet undefined operational environment.

In this regard, the predictors studied can be conveniently divided into two categories. Coders in the first category perform exactly the same operation on both the silent and nonsilent portions of the input signal. In computing the first and second moment statistics used to design these predictors, no distinction is made between speech and silence in the raw data. As a result, the silence-nonsilence ratio for the input data directly influences the value of the predictor coefficients themselves. As a further consequence, it becomes impossible to extrapolate the results observed from these predictors in anything but a qualitative way to speech with a greater percentage of silence.

Coders in the second category operate differently on the two portions of the input, as is discussed in Section IV. In collecting the statistics needed to design and evaluate these predictor models, the speech and silence intervals in the speech sample are handled separately. In this case, the predictor coefficients are (essentially) independent of the input silence-nonsilence ratio. Consequently, one can compute and plot coder performance as a function of this ratio. The only interpolation problem remaining is that of locating an "operating point" on these curves. That is, the problem of determining what the silence-nonsilence ratio would be in an operational situation.

Finally, the limitations imposed by the relatively small number of talkers used in the program should be mentioned. As is always the case throughout speech processing work, one can anticipate experimental results to vary between different talkers and between individual talkers and the "average over all talkers." For the most part, discussion of specific talker-variability effects is scattered throughout this report. At this point a brief comment is offered concerning the intended scope of the study and the attendant viewpoint from which the result must be regarded.

Because information theoretic measures in general have a curious and hard-to-anticipate tendency to suppress some characteristics of ensembles and to enhance others, in establishing a program such as this, it is dangerous to speculate about the probable effect or importance of known characteristics of the source under study. Obtaining crude estimates of the influence of talker-variability has been a consideration in design of the research program. The choice of 10 talkers for the study represents the inevitable compromise between practicality and the definiteness of results. However, in view of the observed variations in output bit rate, it is clear in hindsight, at least, that the number of talkers used does limit the precision of the measurements and reliability with which they can be extended to operational situations.

G. BLOCK DIAGRAM OF AN ANALYSIS OF A PREDICTIVE CODING SYSTEM

At this point, the fundamental theory, language, and methodology of this program should have been made clear. The gap between understanding what is to be done to achieve ideal coding, and what is necessary for an evaluation of such an approach, is illustrated by the analysis block diagram in Figure 5.

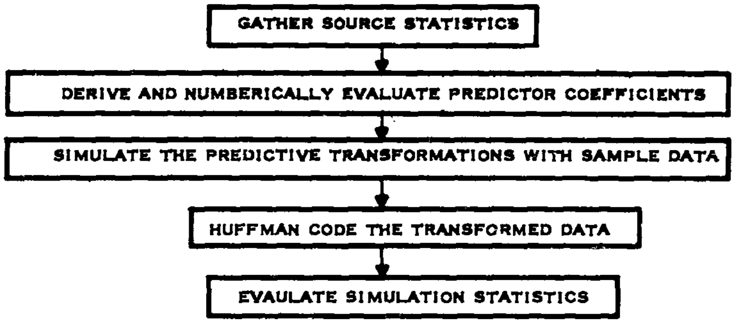

GATHER SOURCE STATISTICS
DERIVE AND NUMBERICALLY EVALUATE PREDICTOR COEFFICIENTS
SIMULATE THE PREDICTIVE TRANSFORMATIONS WITH SAMPLE DATA
HUFFMAN CODE THE TRANSFORMED DATA
EVAULATE SIMULATION STATISTICS
13898

Figure 5. Block Diagram of Analysis for Each Model

Sections III and IV contain a detailed discussion of the analysis implied by this diagram. Section V discusses the conclusions and overall evaluation of results acheived during this investigation.

SECTION III

DERIVATION AND COMPUTATION OF COEFFICIENTS FOR LINEAR PREDICTION

A. MATRIX NOTATION AND REPRESENTATION OF SOURCE STATISTICS

It has been indicated, that for the linear form of predictive transformation,

$$y_i = x_i - p_i \ , \tag{11}$$

several processing models may be considered by the statistical or non-statistical manner of choosing the predictor coefficients. (Recall that x_i represents a message symbol and p_i is a linear prediction of the value of that symbol.) In this section linear predictors derived from known source statistics are discussed. It will assist the analysis of predictor models to first look at the matrix representation of the general linear predictive transformation and the computation of source statistics.

The method of ordering past symbol values in the prediction process is rather arbitrary, but a choice must be made for clarity of analysis. As before, let a subscripted letter x represent a value from the time series being transformed. This time series may be the message consisting of vocoder spectrum amplitude values, or the result of differencing or extracting mean values from the symbols of that message. In either instance, the time series may be described as a vector sequence, each vector representing a set of values for symbols in the vocoder message frame format. With an 18 channel vocoder, there are 18 components for each frame vector. Figure 6 shows a typical part of such a vector sequence, and introduces a subscript notation for reference. (It is important that the notation introduced here be well understood so that subsequent analysis and discussion need not be burdened with detracting explanations.) To avoid the confusion of multiple subscripts, superscripts, and other scripts, an arbitrary multiplexing order is defined for the symbol elements of this time series and the symbols that represent prediction error in the transformation of a current message vector. Corresponding to the message

$$\underset{\text{Past}}{\ldots,} \begin{bmatrix} x_{18,\ t-T} \\ x_{17,\ t-T} \\ \cdot \\ \cdot \\ \cdot \\ x_{1,\ t-T} \end{bmatrix} , \ldots , \begin{bmatrix} x_{18,\ t-1} \\ x_{17,\ t-1} \\ \cdot \\ \cdot \\ \cdot \\ x_{1,\ t-1} \end{bmatrix} , \begin{bmatrix} x_{18,\ t} \\ x_{17,\ t} \\ \cdot \\ \cdot \\ \cdot \\ x_{1,\ t} \end{bmatrix} \underset{\text{Future}}{, \ldots}$$

Figure 6. Sample of T + 1 Frames From a Discrete Vector Time Series

vector at time, t, there is represented a transformed vector following prediction, where the correspondence between the singly and doubly subscripted elements of the vector y is important. Also, the vector y is not explicitly assigned a time reference subscript. The vector representing a multiplexed T + 1 frame sample of message symbols is denoted by,

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_{18} \end{bmatrix} = \begin{bmatrix} y_{18,\,t} \\ y_{17,\,t} \\ \cdot \\ \cdot \\ \cdot \\ y_{1,\,t} \end{bmatrix} = \begin{bmatrix} x_{18,\,t} - p_{18,\,t} \\ x_{17,\,t} - p_{17,\,t} \\ \cdot \\ \cdot \\ \cdot \\ x_{1,\,t} - p_{1,\,t} \end{bmatrix}, \qquad (12)$$

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_{18} \\ x_{19} \\ \cdot \\ \cdot \\ \cdot \\ x_{36} \\ \cdot \\ \cdot \\ \cdot \\ x_{M-17} \\ \cdot \\ \cdot \\ \cdot \\ x_M \end{bmatrix} = \left[\begin{array}{l} \left.\begin{array}{l} x_{18,\,t} \\ x_{17,\,t} \\ \cdot \\ \cdot \\ \cdot \\ x_{1,\,t} \end{array}\right\} \text{"Current" frame} \\ \left.\begin{array}{l} x_{18,\,t-1} \\ \cdot \\ \cdot \\ \cdot \\ x_{1,\,t-1} \end{array}\right\} \text{Immediate past frame} \\ \cdot \\ \cdot \\ \cdot \\ \left.\begin{array}{l} x_{18,\,t-1} \\ \cdot \\ \cdot \\ \cdot \\ x_{1,\,t-T} \end{array}\right\} \text{T}^{\text{th}} \text{ past frame} \end{array}\right] \qquad (13)$$

where, $M = 18 (T + 1)$, is the number of components of the vector x. When discussing a sample or collection of this form, it is occasionally convenient to use the descriptors given at the right of the vector in Equation (13).

Utilizing the above notion, every linear predictive transformation of the type, Equation (11), can be represented in matrix form by,

$$y = Px \tag{14}$$

where P is an 18 x M matrix. The elements of P are the predictor coefficients. This transformation indicates the processing for all 18 symbols of the current frame utilizing the values of symbols from T past frames in the time series. For time-invariant prediction, the elements of P are constants. The numerical values of these elements depend upon the form of the prediction model and the sense of "optimum." Without restricting the form, one can discuss the optimum criteria, and derive a general formulation for use in the succeeding sections.

The least mean square error criterion requires a definition of "expected value." In most applications, expectation is meant in an explicit statistical sense. The general form of the expected value, or mean, of a random variable, v, is given by

$$E(v) = \Sigma vp(v), \tag{15}$$

where p(v) is the probability or probability density associated with the values of v, and the sum is taken over the entire measure space, $\Sigma p(v) = 1$. In many instances, such as the problem at hand, the probability function is not known and sample estimation methods are used to approximate the expectation. In particular, it may be shown that for reasonably large η, the sample estimate

$$E(v) = \frac{1}{\eta} \sum_{i=1}^{\eta} v_i \tag{16}$$

is a meaningful approximation. Note that it is equivalent to Equation (15) with p(v) replaced by the sample estimate of the relative frequency of occurrence of symbol v. When the symbol v in Equation (16) represents a variable from a discrete time series, the sum is taken over a set of observations of this variable from a sample of the series. This measure of expectation is also appropriate for estimates of the second, or product, moments that are necessary for deriving LLS predictor coefficients. The second moments may be formed about an arbitrary value of the variable. With the above definition, the expected second moment of a variable will be a minimum when computed about the mean, and similarly for product moments of two such variables.

The expected value of a random vector (random variables as component elements) is defined to be the vector of expected values for the components. The mean value of the vector, x, is thus given by,

$$\bar{x} = E[x] = \begin{bmatrix} E\,[x_I] \\ E\,[x_2] \\ \cdot \\ \cdot \\ \cdot \\ E\,[x_M] \end{bmatrix} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \cdot \\ \cdot \\ \cdot \\ \bar{x}_M \end{bmatrix} . \tag{17}$$

Two second moments for the components of the vector x are of interest. The value $E(x_i x_i)$ is the estimate of the variance of the ith component element, x_i. The value $E(x_i x_j)$ is the estimate of the covariance between the ith and jth component elements. The matrix composed of such estimates is called the sample covariance matrix, and is conveniently denoted by

$$\Sigma = E[xx'] = \{\sigma_{ij}\} = \{E(x_i x_j)\} , \tag{18}$$

where the prime denotes transposition. Since the expectation of a product is independent of the order of factors, $\sigma_{ij} = \sigma_{ji}$, and Σ is a symmetric matrix. In the above notation, the moments are implied to be about the origin. For moments about the mean,

$$\begin{aligned} E[(x_i - \bar{x}_i)(x_j - \bar{x}_j)] &= E[x_i x_j - \bar{x}_i x_j - x_i \bar{x}_j \; \bar{x}_i x_j] \\ &= E[x_i x_j] - \bar{x}_i \bar{x}_j , \end{aligned} \tag{19}$$

since the expectation operation is linear. Note that for a random vector x defined by $x = v - \bar{v}$, where v is an arbitrary random vector, it follows that,

$$\bar{x} = E[x] = E[v - \bar{v}] = \bar{v} - \bar{v} = 0 \text{ (vector)}. \tag{20}$$

It is assumed in what follows that the message sample vector, x, is such a vector with mean removed.

The expected value of a constant matrix times a random vector (linear transformation) is the constant matrix times the expected value of the vector. That is, for $y = Mx$,

$$\bar{y} = E[y] = E[Mx] = ME[x] = M\bar{x}. \tag{21}$$

Thus, for the general linear predictive transformation, Equation (14), the covariance matrices of the message time series before and after transformation are related by

$$S = E[yy'] = E[Pxx'P'] = PE[xx'] P' = P\Sigma P', \quad (22)$$

where Σ is defined by Equation (18) and s_{ij} will denote the typical element of the matrix S. The diagonal elements, s_{ii}, of S are the variances of the transformed symbol series, which are also the expected squared error of the prediction scheme. All variance-covariance estimates are proportional to the linear correlation between the respective variables. The application of the minimum mean square error criterion implies that one chooses the elements of P, subject to other constraints on the form of predictors, such that the diagonal elements are a minimum. It will follow, that the off-diagonal (covariance) elements have expected value zero. Thus, the linear correlation of intersymbol influence is expected to be zero for the transformed sequence, implying linear independence of the symbols.

One subtle feature needs to be clarified in the discussion of computing expected values. To compute a sample covariance matrix such as Σ, an assumption of stationarity is made regarding the nature of the message source. If the time index of such a sample is indicated for the moment by a superscript, the set of vectors $(x^{\tau}, \tau = \tau_1, \tau_2, \dots \tau_{\eta})$, may be described as η sample vectors taken from an appropriate speech sample at frame times denoted by τ. With this notation, the sample covariance matrix is given by,

$$\Sigma = \frac{1}{\eta} \sum_{\tau=\tau_1}^{\tau_\eta} x^{\tau}(x^{\tau})' \quad (23)$$

or

$$\sigma_{ij} = \frac{1}{\eta} \sum_{\tau=\tau_1}^{\tau_\eta} x_i^{\tau} x_j^{\tau}, \quad \begin{array}{l} i = 1, 2, \dots, M \\ j = 1, 2, \dots, M. \end{array} \quad (24)$$

One further notes that the multiplexing of frame data to form the vectors (x^{τ}) implies a relation between the component subscript and the time index. For instance, the first component of a vector x^{τ} at time t is the same as the 19th component of the vector x^{τ} taken at a time one frame later. Any assumption regarding the lack of dependence of the sample vectors on the time index is an assumption of stationarity of the source ensemble. Obviously, for the sample covariance matrix to be independent of time, some assumption of stationarity must be made and the matrix computed accordingly. It has been assumed that expected values of the source statistics are independent of the frame time unit used in Equation (24). This assumption is reasonable, whereas, an assumption that the multiplexing could be accomplished with a symbol time index, and statistics computed accordingly, was evaluated and found quite unreasonable. We emphasize that the assumption of stationarity over frame time units is the primary motivating factor for use of matrix representation.

A familiarity with the above notation, and an understanding of what and how source statistics are accumulated, are sufficient background for developing explicit prediction forms.

B. EQUAL LENGTH PREDICTORS (ELP)

The general form of linear predictive transformation was introduced as Equation (14). In this section the system of equations is derived that must be solved for the coefficients of predictors that are separate and of fixed length for each vocoder spectrum channel. This form will be compared in the next section with a model where the length of the predictors are not the same for each channel. The reasons for developing both models will be discussed at that point.

The equal length predictor model is denoted by the matrix transformation,

$$y = Ax,$$

where,

$$A = (a_{ij}) = \begin{bmatrix} 1 & \alpha_{11} & \cdot & \cdot & \cdot & \alpha_{116} & \alpha_{117} & \cdot & \cdot & \cdot & \alpha_{1N} & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 1 & \alpha_{21} & \cdot & \cdot & \cdot & \alpha_{216} & \alpha_{217} & \cdot & \cdot & \cdot & \alpha_{2N} & 0 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & 0 & 1 & \alpha_{171} & \cdot & \cdot & \cdot & \alpha_{17N-16} & \alpha_{17N-15} & \cdot & \cdot & \cdot & \alpha_{17N} & 0 \\ 0 & 0 & \cdot & \cdot & \cdot & 0 & 1 & \alpha_{181} & \cdot & \cdot & \cdot & \alpha_{18N-16} & \alpha_{18N-15} & \cdot & \cdot & \cdot & \alpha_{18N} \end{bmatrix} \qquad (26)$$

and the vectors y and x are as before. The order of elements in the numerical computation of predictor coefficients is a confusing permutation of the order used to indicate the computation envolved in the processing of actual messages. The issue is further complicated by the constraints of the model which are represented by the correspondence of a typical element, a_{ij}, with the appropriate element as given explicitly in Equation (26). This correspondence is further illustrated by the equations,

$$y_i = \sum_{j=1}^{M} a_{ij} x_j \, , \qquad (27)$$

$$= \sum_{j=i+1}^{i+N} a_{ij} x_j, \qquad (28)$$

$$= x_i + \sum_{k=1}^{N} \alpha_{ik} x_{i+k}, \qquad (29)$$

where the latter equation will be used in the derivation because it has the constraints of the prediction form "built-in", and is more anologous to Equation (11). The change of sign preceeding the summation is an algebraic convenience in no way restricting the generality, since the numerical values of the coefficients are not restricted in sign.

The length of the predictors, N, have been restricted to be a multiple of 18. This restriction is not necessary, since N could be arbitrary. For this analysis, however, the computations are more efficient when working with an integral multiple of sample frames. Also, with the length restricted in this manner, more consistant prediction results for each symbol of the transformed message frame. Note that the form of Equation (26) implies that M = N + 18. Or, stated another way, M = 18 (T + 1) = N + 18, implies that message sample frames such that N = 18T are used for prediction and for computation of necessary source statistics.

The optimum rms predictor coefficients are those which minimize the diagonal elements of the matrix S, Equation (20). With the notation of Equation (29),

$$s_{ii} = E\,[y_i y_i]$$

$$= E[x_i^2 + 2x_i \sum_{k=1}^{N} \alpha_{ik} x_{i+k} + (\sum_{k=1}^{N} \alpha_{ik} x_{i+k})^2] . \qquad (30)$$

To avoid confusion between the sample convariance matrix, Σ, and a different matrix composed of elements from Σ, an alternate notation is introduced for the expected second moments, (recalling also that the expected mean is zero).

$$\phi_{u,v} = \phi_{v,u} = E(x_u x_v) = \sigma_{uv} = \sigma_{vu}. \qquad (31)$$

It follows from Equation (30) that,

$$s_{ii} = \phi_{i,i} + 2 \sum_{k=1}^{N} \alpha_{ik} \phi_{i,i+k} + 2 \sum_{k=1}^{N} \sum_{\ell=1}^{N} \alpha_{ik} \alpha_{i\ell} \phi_{i+\ell,i+k}. \qquad (32)$$

The necessary conditions for a minimum over the set of undetermined coefficients are given by,

$$\frac{\partial s_{ii}}{\partial \alpha_{im}} = \phi_{i,\ i+m} + \sum_{k=1}^{N} \alpha_{ik}\,\phi_{i+m,\ i+k} = 0, \quad \begin{array}{l} i = 1, 2, \ldots, 18 \\ m = 1, 2, \ldots, N \end{array} . \qquad (33)$$

These conditions specify the 18 N x N systems of linear equations,

$$\Phi_i p_i + C_i = 0, \quad i = 1,\ 2,\ \ldots,\ 18, \qquad (34)$$

or

$$\begin{bmatrix} \phi_{i+1,\,i+1} & \phi_{i+1,\,i+2} & \cdot & \cdot & \cdot & \phi_{i+1,\,i+N} \\ \phi_{i+2,\,i+1} & \phi_{i+2,\,i+2} & \cdot & \cdot & \cdot & \cdot \\ \phi_{i+3,\,i+1} & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \phi_{i+N,\,i+1} & \cdot & \cdot & \cdot & \cdot & \phi_{i+N,\,i+N} \end{bmatrix} \begin{bmatrix} \alpha_{i1} \\ \alpha_{i2} \\ \alpha_{i3} \\ \cdot \\ \cdot \\ \cdot \\ \alpha_{iN} \end{bmatrix} = - \begin{bmatrix} \phi_{i+1,\,i} \\ \phi_{i+2,\,i} \\ \phi_{i+3,\,i} \\ \cdot \\ \cdot \\ \cdot \\ \phi_{i+N,\,i} \end{bmatrix} ,$$

for i = 1, 2, ..., 18. The solutions of these systems are the optimum predictor coefficients. It is apparent from inspection of Φ_i, that for each appropriate i, Φ_i is a (major diagonal) principle submatrix of the covariance matrix, Σ. Further, all elements of Σ, except σ_{11}, are utilized at one point or another in the system of Equation (34). For each system, the constant vector, c_i, is a bordering row or column of Φ_i superimposed on Σ. These and other features of interest are discussed in more detail in Section III. D, where the numerical algorithm for solution of Equation (34) is developed.

It is worthwhile to develop the expressions for the expected minimum variance and expected covariance of the transformed symbols, that is, subject to the solutions of Equation (34). The general expression for s_{ij} is given by,

$$s_{ij} = \phi_{i,\,j} + \sum_{\ell=1}^{N} \alpha_{j\ell}\,\phi_{i,\,j+\ell} + \sum_{k=1}^{N} \alpha_{ik}\,\phi_{j,\,i+k} + \sum_{k=1}^{N} \sum_{\ell=1}^{N} \alpha_{ik}\,\alpha_{j\ell}\,\phi_{i+k,\,j+\ell} . \qquad (35)$$

Now, for i = j, and using the conditions Equation (33),

$$
\begin{aligned}
s_{ii} &= \phi_{ii} + 2 \sum_{k=1}^{N} \alpha_{ik} \phi_{i,i+k} + \sum_{\ell=1}^{N} \alpha_{i\ell} \left\{ \sum_{k=1}^{N} \alpha_{ik} \phi_{i+k,i+\ell} \right\} \\
&= \phi_{ii} + 2 \sum_{k=1}^{N} \alpha_{ik} \phi_{i,i+k} + \sum_{\ell=1}^{N} \alpha_{i\ell} \left\{ -\phi_{i,i+\ell} \right\} \\
&= \phi_{ii} + \sum_{k=1}^{N} \alpha_{ik} \phi_{i,i+k}, \qquad i = 1, 2, \ldots, 18.
\end{aligned} \tag{36}
$$

This expression gives the value of the theoretical minimum variance of the transformed variables. Theoretical, in the sense that it is an estimate based on the sample statistics used in the computation of the optimum predictor coefficients. This estimate may conveniently be computed along with the numerical computation of the coefficients, thus giving a value to compare with the ϕ_{ii} values, the sample variance of the original variables. An estimate of the symbol and/or frame entropy will be developed in terms of the theoretical variances.

For i ≠ j, noting again the symmetry of the ϕ_{uv} elements, and with the results of Equation (33), Equation (35) becomes,

$$
\begin{aligned}
s_{ij} &= \phi_{ij} + \sum_{\ell=1}^{N} \alpha_{j\ell} \phi_{i,j+\ell} + \sum_{k=1}^{N} \alpha_{ik} \phi_{j,i+k} + \sum_{\ell=1}^{N} \alpha_{j\ell} \left\{ -\phi_{i,j+\ell} \right\} \\
&= \phi_{ij} + \sum_{k=1}^{N} \alpha_{ik} \phi_{j,i+k} \\
&= 0
\end{aligned} \tag{37}
$$

since, with no loss of generality, let j = i + m and the results of Equation (33) provide the last step. This result verifies the previous statement that the expected linear correlation between symbols is zero for the transformed symbol series.

Before considering the numerical problems associated with solving the system of Equation (34), the battle of subscripts is continued for a development paralleling that above, but for a predictor model of different form.

C. UNEQUAL LENGTH PREDICTORS (ULP)

In terms of the general linear predictive transformation, Equation (14), the model to be considered here is denoted by,

$$
y = Bx, \tag{38}
$$

where,

$$B = (b_{ij}) = \begin{bmatrix} 1 & \beta_{12} & \beta_{13} & \cdot & \cdot\;\cdot\;\cdot & & \cdot & \cdot & \cdot\;\cdot\;\cdot & \beta_{1M} \\ 0 & 1 & \beta_{23} & \beta_{24} & \cdot\;\cdot\;\cdot & & \cdot & \cdot & \cdot\;\cdot\;\cdot & \beta_{2M} \\ \cdot\;\cdot & \cdot & \cdot & \cdot & \cdot\;\cdot\;\cdot & & \cdot & \cdot & \cdot\;\cdot\;\cdot & \cdot \\ \cdot\;\cdot & \cdot & \cdot & \cdot & \cdot\;\cdot\;\cdot & & \cdot & \cdot & \cdot\;\cdot\;\cdot & \cdot \\ \cdot\;\cdot & \cdot & \cdot & \cdot & \cdot\;\cdot\;\cdot & & \cdot & \cdot & \cdot\;\cdot\;\cdot & \cdot \\ 0\;\cdot & \cdot & \cdot & 0 & 1 & \beta_{1718} & \beta_{1719} & \cdot & \cdot\;\cdot\;\cdot & \beta_{17M} \\ 0\;\cdot & \cdot & \cdot & 0 & 0 & 1 & \beta_{1819} & \beta_{1820} & \cdot\;\cdot\;\cdot & \beta_{18M} \end{bmatrix} \tag{39}$$

As before, and for the same reasons, the correspondence between the elements, b_{ij}, and the constraints of the form of prediction are given by Equation (39) and the equations,

$$\left.\begin{aligned} y_i &= \sum_{j=1}^{M} b_{ij}x_j && (40) \\ &= x_i + \sum_{k=i+1}^{M} \beta_{ik}x_k && (41) \end{aligned}\right\} \quad i = 1, 2, \dots, 18.$$

The change in coefficient subscript notation is typical of the change in emphasis between this model and the previous. With ELP, the emphasis was placed on utilizing a fixed number of past symbols for prediction. With ULP, the emphasis is placed on using all symbols from a fixed number of past frames. Both models have merit as a prediction scheme. The significant difference in the numerical computation of coefficients motivated the investigation of both models.

When processing with the ULP model, the number of past symbols utilized for prediction varies with the ordering of symbols in the frame. For instance, if no past frames are used, the first symbol of a current frame must be processed without prediction. The second symbol is processed utilizing the known value of the first, etc., until the 18th symbol is processed with prediction based upon the known first 17 symbol values. This limited type of prediction is called "across channel" prediction. With speech sources, across channel prediction would be expected to remove intersymbol influence characterized by the format structure of voiced sounds. When past frame values are utilized in the prediction, predictor lengths are increased by a constant multiple of 18, and the influence of frame-to-frame (time) correlation is hopefully removed by the linear approximation. Obviously, the ELP model uses the same source correlation statistics and includes across channel and back in time prediction. For long (i. e., utilizing several past frames) predictor lenghts the two models would be expected to give about the same results. This conclusion has been verified.

The derivation of the systems of equations for the rms optimum ULP coefficients proceeds much the same as that for the ELP model. With the notation,

$$\theta_{u,v} = \theta_{v,u} = E(x_u x_v) = \sigma_{uv} = \sigma_{vu}, \tag{42}$$

the sample covariance expression is,

$$s_{ij} = \theta_{i,j} + \sum_{k=i+1}^{M} \beta_{ik}\,\theta_{k,j} + \sum_{\ell=j+1}^{M} \beta_{j\ell}\,\theta_{\ell,j} + \sum_{k=i+1}^{M}\ \sum_{\ell=j+1}^{M} \beta_{ik}\,\beta_{j\ell}\,\theta_{k,\ell}, \tag{43}$$

for i, j = 1, 2, . . . , 18. The expression for s_{ii} reduces to,

$$s_{ii} = \theta_{ii}1 + 2 \sum_{k=i+1}^{M} \beta_{ik}\,\theta_{k,i} + \sum_{k=i+1}^{M}\ \sum_{\ell=i+1}^{M} \beta_{ik}\,\beta_{i\ell}\,\theta_{k,\ell}, \tag{44}$$

and the minimization equations are,

$$\frac{\partial s_{ii}}{\partial \beta_{im}} = 0, \qquad \begin{array}{l} i = 1, 2, \ldots, 18 \\ m = i+1,\ i+2, \ldots, M^{\cdot} \end{array} \tag{45}$$

These conditions imply the systems,

$$\sum_{k=i+1}^{M} \beta_{ik}\,\theta_{m,k} + \theta_{m,i} = 0, \tag{46}$$

for i = 1, 2, . . . , 18, and m = i + 1, i + 2, . . . , M; or, equivalently, the matrix systems,

$$\begin{bmatrix} \theta_{i+1,\,i+1} & \theta_{i+1,\,i+2} & \cdot & \cdot & \cdot & \theta_{i+1,\,M} \\ \theta_{i+2,\,i+1} & \theta_{i+2,\,i+2} & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \theta_{M,\,i+1} & \cdot & \cdot & \cdot & \cdot & \theta_{M,\,M} \end{bmatrix} \begin{bmatrix} \beta_{i\ i+1} \\ \beta_{i\ i+2} \\ \cdot \\ \cdot \\ \cdot \\ \beta_{iM} \end{bmatrix} = - \begin{bmatrix} \theta_{i+1,\,i} \\ \theta_{i+2,\,i} \\ \cdot \\ \cdot \\ \cdot \\ \theta_{M,\,i} \end{bmatrix}, \tag{47}$$

for i = 1, 2, ..., 18. Again, the source statistics represented by Σ are sufficient to define the optimum predictor coefficients with the rms criterion.

Comparing Equations (34) and (47), at least one point of similarity is apparent. For N = M − 18 = 18T, the two systems have the same solution for i = 18. That is, the optimum predictor of length N is the same for each model. Omitting the restriction that N be a multiple of 18, it is straight forward to show that all optimum ULP predictors are identical to the equivalent channel predictor in an ELP model of appropriate length.

Any obvious similarity ends there. The more significant comparisons are made in the framework of the numerical computation algorithm used to solve the two types of systems. Before proceeding to discuss that algorithm, we finish the parallel to the analysis of ELP.

Introducing the results of Equation (46) into the general sample covariance estimate for ULP, `Equation (43) we have

$$
\begin{aligned}
s_{ij} &= \theta_{ij} + \sum_{k=i+1}^{M} \beta_{ik}\, \theta_{k,j} + \sum_{\ell=j+1}^{M} \beta_{j\ell}\, \theta_{\ell,i} + \sum_{\ell=j+1}^{M} \beta_{j\ell} \left\{ \sum_{k=i+1}^{M} \beta_{ik}\, \theta_{k,\ell} \right\} \\
&= \theta_{ij} + \sum_{k=i+1}^{M} \beta_{ik}\, \theta_{k,j} + \sum_{\ell=j+1}^{M} \beta_{j\ell}\, \theta_{\ell,i} + \sum_{\ell=j+1}^{M} \beta_{j\ell} \left\{ -\theta_{\ell,i} \right\} \\
&= \theta_{ij} + \sum_{k=i+1}^{M} \beta_{ik}\, \theta_{k,j} \quad , \ i, j = 1, 2, \ldots 18. \qquad (48)
\end{aligned}
$$

Now, for i ≠ j, let j = i + m with no loss of generality. It follows from Equation (46) that $s_{ij} = 0$, i ≠ j, as with the ELP model. For i = j, the theoretical expected minimum variance is,

$$
s_{ii} = \theta_{ii} + \sum_{k=i+1}^{M} \beta_{ik}\, \theta_{k,i} \quad , \ i = 1, 2, \ldots, 18, \qquad (49)
$$

anologous to the ELP result, Equation (36).

D. COMPUTATIONAL ALGORITHMS FOR SOLUTION OF ELP AND ULP COEFFICIENTS

The computational techniques for solving systems of Equations (34) and (47) are too closely associated with the merits of each model to be relegated to a reference to the mathematical literature. In fact, the systems representing the desired solution for the ULP model are the major portion of the little-known "Escalator Method"[7] for inversion of matrices by a bordering scheme. The advantages of this numerical procedure and its match with the ULP problem are just what is required to efficiently examine the prediction scheme for long sequences. A minimum development of this computational algorithm is given here to indicate the difficulties and importance of computing accurate predictor coefficients.

The fundamental role in solving linear systems of equations is the inversion, explicitly or implicitly, of the coefficient matrix. The basis of any bordering method to accomplish that task is given in terms of the relations,

$$A_n = \begin{bmatrix} A_{n-1} & u_n \\ v_n & a_{nn} \end{bmatrix} , \qquad (50)$$

and,

$$A_n^{-1} = \begin{bmatrix} A_{n-1}^{-1} + \dfrac{A_{n-1}^{-1} u_n v_n A_{n-1}^{-1}}{a_{nn} - v_n A_{n-1}^{-1} u_n} & - \dfrac{A_{n-1}^{-1} u_n}{a_{nn} - v_n A_{n-1}^{-1} u_n} \\ - \dfrac{v_n A_{n-1}^{-1}}{a_{nn} - v_n A_{n-1}^{-1} u_n} & \dfrac{1}{a_{nn} - v_n A_{n-1}^{-1} u_n} \end{bmatrix}, \qquad (51)$$

where A_n is an n x n matrix partitioned in a bordering manner as indicated in Equation (50). For a symmetric matrix, the row vector, v_n, is equal to the transpose of the column vector u_n. Equation (51) gives the formula for the inverse of an n x n matrix in terms of the inverse of the (n - 1) x (n - 1) major submatrix and the bordering elements. Starting with the inverse of any order submatrix (even n = 2, the scalar, $A_{n-1} = a_{11}$, with $A_{n-1}^{-1} = 1/a_{11}$) of a matrix, $A_{N,N}$, a "march" type order of computation may be used to compute the inverse of higher order matrices, until A_{NN}^{-1} is finally evaluated. Only the simple algebraic operations are required at each step in this direct method using Equation (51).

This bordering method is applied to the solution of a symmetric linear system of equations in the following manner. Let the system at the k^th stage of representation be written as,

$$A_k X_k = F_k, \qquad (52)$$

with the notation,

$$A_k = \begin{bmatrix} A_{k-1} & u_k \\ u_k & a_{kk} \end{bmatrix}, F_n = \begin{bmatrix} F_{k-1} \\ f_k \end{bmatrix}, \text{ and } X_k = \begin{bmatrix} y \\ x_k \end{bmatrix} . \qquad (53)$$

Writing the solution of Equation (52) as,

$$X_k = A_k^{-1} F_k, \tag{54}$$

and incorporating the formula of Equation (51), we have,

$$X_k = \begin{bmatrix} y \\ \\ x_k \end{bmatrix} - \begin{bmatrix} A_{k-1}^{-1} F_{k-1} \\ \\ 0 \end{bmatrix} + \frac{1}{a_{kk} - u_k' A_{k-1}^{-1} u_k} \begin{bmatrix} A_{k-1}^{-1} u_k u_k' A_{k-1}^{-1} F_{k-1}^{-1} - A_{k-1}^{-1} u_k f_k \\ \\ - u_k' A_{k-1}^{-1} F_{k-1} + f_k \end{bmatrix} . \tag{55}$$

To simplify the notation, let $A_{k-1}^{-1} \; F_{k-1} = X_{k-1}$, and $A_{k-1}^{-1} u_k = -Z_{k-1}$. Then,

$$X_k = \begin{bmatrix} y \\ \\ x_k \end{bmatrix} = \begin{bmatrix} X_{k-1} \\ \\ 0 \end{bmatrix} + \frac{f_k - u_k' X_{k-1}}{a_{kk} + u_k' Z_{k-1}} \begin{bmatrix} Z_{k-1} \\ \\ 1 \end{bmatrix} . \tag{56}$$

Now in the general application, F_k is usually extended at each stage so that X_{k-1} is the solution of the previous stage, and X_N is the desired solution of the final stage, $A_{NN} X_N = F_N$. This leaves the determination of Z_{k-1} as the major effort at each stage of the march. If A_{k-1}^{-1} is computed by Equation (51) explicitly and saved at each step, Z_{k-1} is formed by simple multiplication. However, for large systems, there are better ways of computing Z_{k-1} at each stage.

Now, without showing all details, the following remarks are made concerning the general bordering method and the ULP system of equations given by Equation (47):

(1) The ULP systems for all $M_\tau = 18 (\tau + 1)$, $\tau = 0, 1, \ldots, T$, may be combined as one matrix model. That is, if y^τ denotes the transformed frame vector at any time which utilizes symbol values from τ past frames, following one has, Equation (57), as a generalized form of Equation (38). All the development of Paragraph C applies directly to this generalized transformation. Each horizontal partitioning of 18 rows is a system as discussed above. These T + 1 systems are merely "stacked" together to more easily demonstrate that solutions to all lesser systems are by-products of the solution of the higher order system.

$$\begin{bmatrix} y_1^T \\ y_2^T \\ \cdot \\ \cdot \\ \cdot \\ y_{18}^T \\ y_1^{T-1} \\ \cdot \\ \cdot \\ \cdot \\ y_{18}^{T-1} \\ \cdot \\ \cdot \\ \cdot \\ y_1^o \\ \cdot \\ \cdot \\ \cdot \\ y_{18}^o \end{bmatrix} = \begin{bmatrix} 1 & \beta_{12} & \beta_{13} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \beta_{1M} \\ 0 & 1 & \beta_{23} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \beta_{2M} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & 0 & 1 & \beta_{1819} & \cdot & \cdot & \cdot & \cdot & \cdot & \beta_{18M} \\ 0 & \cdot & \cdot & \cdot & 0 & 1 & \beta_{1920} & \cdot & \cdot & \cdot & \cdot & \beta_{19M} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & 0 & 1 & \beta_{3637} & \cdot & \cdot & \beta_{36M} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 & 1 & \beta_{M-17\,M-16} & \beta_{M-17\,M} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_{18} \\ x_{19} \\ \cdot \\ \cdot \\ \cdot \\ x_{36} \\ \cdot \\ \cdot \\ \cdot \\ x_{M-17} \\ \cdot \\ \cdot \\ \cdot \\ x_M \end{bmatrix}, \qquad (57)$$

(2) By "Flipping"* the system of Equation (57) and systems of the form of Equation (47), the solutions for the ULP coefficients are given as the Z_{k-1} vectors in a bordering matrix inversion algorithm for the solution of the longest vector. That is, at each stage of the march, one solution vector (a row of Equation (57)) is determined. All solution vectors of length less than N are determined as by-products of the algorithm for solving the N^{th} order system. The a_{ij} elements of the typical bordering system illustrated in this section, correspond to the elements of the "flipped" Σ matrix of previous sections.

(3) The theoretical variance of the transformation variable corresponding to a predictor of length k – 1 is given by the $a_{kk} + u_k^{'} Z_{k-1}$ expression at each stage of the march as given by Equation (56). That is, the theoretical variance at each stage requires no special computation.

* "Flipping" means an inversion of order of all elements in each row and column of the matrices.

(4) This algorithm is extremely stable for computations with high order systems. Numerical round-off errors do not accumulate, and a convenient check on numerical error is easily formed at each stage of the march.

The conclusion to draw from these remarks is summarized here for clarity. For an ULP model with longest predictor of length, M, an "optimum" computing algorithm exists for determining the coefficients of that predictor. The solution for all M - 1 shorter length predictors of similar ULP models and the theoretical estimates of the minimum variance for all predictors are computed without extra computation in the process. The algorithm is numerically stable for large systems even when such systems are "ill-conditioned." All of these results are achieved at the expense of effort equivalent to solving for one of the 18 predictors of an M-length ELP model.

As one would assume at this point in the presentation, the ULP model was used as an initial approach to LLS prediction. The computation of all predictors indicated in Equation (57) for M = 162 (eight frames back in time), and the corresponding minimum variance estimates, were computed in about 15 minutes per speaker sample. The analysis of these ULP models demonstrated that most of the predictable signal structure that could be removed by LLS prediction was from the across channel and past two or three frames data. These results compared favorably with less accurate estimates that had been formed on the basis of analysis of the multiplexed correlation matrix derived from an approximate M x M Σ matrix.

For the models utilizing prediction no more than two frames back (total of three frames) in time, the ELP model was used. This model gives more consistant prediction across an entire frame for the shorter lengths. The 18 separate N x N systems for each speaker sample were solved using the bordering matrix algorithm. The advantages of the algorithm in this case were simply the stability and speed of computation. No shorter solution vector by-products were analyzed although the algorithm would have facilitated such analysis. For any N = 54, ELP model, all 18 54 x 54 systems can be solved in about five minutes. These time figures are for Fortran Object Programs executed within the Executive (FAST) System for the IBM 7074 digital computer. The most significant features pertaining to the first two blocks in the analysis diagram, Figure 5, have been discussed.

E. ESTIMATION OF APPARENT ENTROPY FROM THEORETICAL VARIANCE ESTIMATES

As mentioned in Paragraph D, theoretical estimates of the minimum variance were computed along with the predictor coefficients for each channel. Any method of attaching significance to these estimates is less than straight forward. The method outlined here has proven useful — the pragmatic qualification.

The transformations derived under the ELP and ULP models with the minimum rms error criterion give predictor coefficients distributed on the real line. Regardless of the analog or discrete nature of the x symbol values, the y symbol values are continuously distributed over some finite range about the expected zero mean value. It is difficult, and for the most part meaningless, to attempt to estimate the entropy of some bounding continuous distribution of values over the range of the y values. It is difficult because the variational extreme problems require numerical, rather than analytical techniques to specify the entropy of the worst case solutions. Such solutions are more academic than meaningful, since the manner of channel coding requires a discrete symbol alphabet; thus, a quantizer is fundamental to the actual processing design. Much more useful results are possible when the theoretical entropy estimates are based upon a discrete distribution obtained by quantization of a gaussian approximation to the unknown continuous distribution.

The approximate symmetry about the zero mean of the distribution of transformed symbol values is one of the very desirable features of the predictive transformations. Furthermore, the symbols from the separate vocoder channels are expected to be linearly independent. These two characteristics imply that an entropy estimate based upon a quantized gaussian approximation to the first order marginal probability distribution is worthy of consideration. The theoretical variance estimates for each channel are sufficient to specify a normal probability distribution about the expected zero mean value. Since the probability associated with a quantization interval is a functional evaluated by integration of the continuous distribution over that interval, the effect is a smoothing of errors of approximation. That is, the probability functional is a better approximation to the unknown true value than the gaussian distribution is an approximation to the unknown true distribution at any given point. Further, when the entropy is estimated from the probabilities associated with the quantization levels, another smoothing of the same nature takes place. These considerations provide heuristic support for the quantized-gaussian approximation model.

Figure 7 shows the method of computing interval probability estimates from the Gaussian approximation. The curve is typical of that for any channel, or for the combined channel series. An arbitrary number of quantization intervals may be taken, usually an odd number and symmetric about zero unless the variance is exceedingly small. The probabilities associated with the bounded intervals are given by

$$P_i = \int_{y_{i-1}}^{y_i} N(y)\, dy, \tag{58}$$

and the intervals at each end are chosen as the infinite tails of the distribution. Of course, the sum of all these discrete probabilities is unity. Thus, a probability is associated with the occurance of symbol values at each level of the quantization alphabet. The average information rate, (entropy), for symbols distributed as this alphabet is given by

$$H(y) = - \sum_i p_i \log p_i . \tag{59}$$

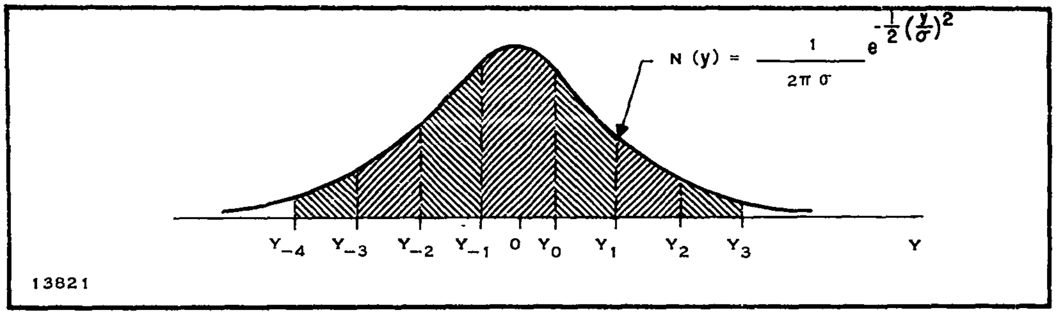

N (y) = $\frac{1}{2\pi\sigma} e^{-\frac{1}{2}(\frac{y}{\sigma})^2}$
Y_{-4}
Y_{-3}
Y_{-2}
Y_{-1}
0
Y_0
Y_1
Y_2
Y_3
Y
13821

Figure 7. Quantized Gaussian Distribution

The procedure just outlined has been shown to give quite accurate estimates, except for σ^2 much less than unity. As the variance decreases below unity, the entropy estimate becomes conservative. For such highly peaked distributions, the gaussian approximation is not accurate. However, the bias appears to be quite consistant for symmetric distributions, and could probably be taken into consideration if accurate estimates were desired for extremely peaked distributions.

Figure 8 is an example plot of entropy versus theoretical variance, based on a seven level quantized normal distribution such as described above. Table II is a typical compilation of theoretical estimates of the minimum variance and approximate entropy for ULP predictors utilizing zero, one, ... eight past frames data. The variance estimates were computed during the computation of predictor coefficients. The entropy estimates were taken from the curve of Figure 8.

The conclusion that removable (predictable) signal structure is primarily evidenced in the first couple of past frames was reached on the basis of results similar to that shown in Table II. Data of this type was prepared for several speaker samples and VAP, V/UV options. Note that (refer to Paragraph C) results vary with order of multiplexing symbols within frames for the ULP model. This is not true for the ELP model.

As a final comment on these estimation procedures, it should be noted again that the effect of quantization noise upon the results cannot be taken into account in the theoretical derivation of the optimum predictors. One would expect that this additive noise in the prediction processing would effect the estimates of reduction in entropy. Further, since the LLS prediction was based upon assumptions of stationary source statistics, there is the question of how well these theoretical estimates will relate to processing of sample data other than that used to accumulate the sample covariance matrix. The simulation models to be described in Section IV were used to provide more information relating to these questions. Simulation results showed that the theoretical estimates, formed as outlined above, were accurate to within 5 percent as the worst case.

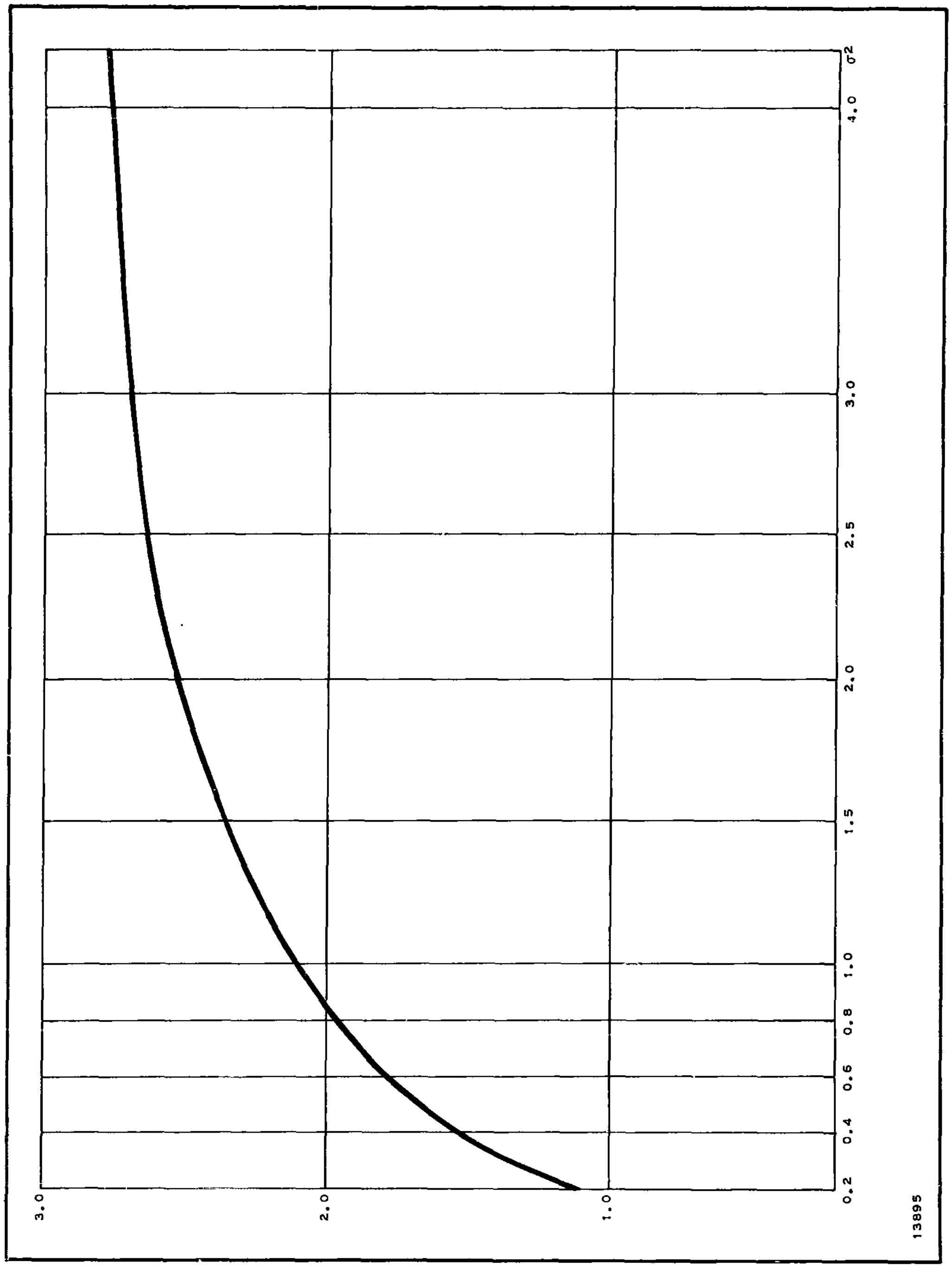


Figure 8. Entropy of Seven-Level Quantized Normal Distribution Versus σ^2 of $N(0, \sigma^2)$

Table II. ULP Model–Theoretical Entropy Estimates

0006 V0002 P06A–Speech Only–All Voice–Reversed Order–1227 Frames

Channel	σ_s^2	σ_0^2	H_0	σ_{-1}^2	H_{-1}	σ_{-2}^2	H_{-2}	σ_{-3}^2	H_{-3}	σ_{-4}^2	H_{-4}	σ_{-5}^2	H_{-5}	σ_{-6}^2	H_{-6}	σ_{-7}^2	H_{-7}	σ_{-8}^2	H_{-8}
1	7.24	3.15	2.70	1.13	2.18	1.08	2.15											0.97	2.08
2	4.59	2.20	2.58	1.07	2.14	1.02	2.11											0.88	2.02
3	6.42	3.19	2.70	1.48	2.35	1.35	2.29											1.23	2.23
4	5.77	1.54	2.37	0.97	2.68	0.93	2.05											0.86	2.00
5	5.89	1.83	2.48	1.08	2.15	1.02	2.11											0.91	2.04
6	3.86	1.48	2.35	0.85	2.00	0.82	1.97											0.74	1.91
7	4.31	1.86	2.48	1.15	2.19	1.06	2.14											0.96	2.07
8	5.72	2.61	2.64	1.40	2.31	1.32	2.28											1.22	2.23
9	4.89	2.42	2.62	1.45	2.34	1.39	2.31											1.28	2.26
10	6.15	3.63	2.72	2.12	2.56	1.98	2.52	●		●		●		●		●		1.81	2.47
11	7.01	4.18	2.77	2.30	2.59	2.15	2.56											1.90	2.49
12	6.69	3.31	2.71	1.95	2.51	1.81	2.47											1.63	2.40
13	6.16	3.81	2.75	1.68	2.42	1.58	2.39											1.39	2.31
14	7.25	5.01	3.00	2.00	2.52	1.86	2.48											1.66	2.42
15	4.90	0.45	1.60	0.36	1.46	0.35	1.43											0.31	1.35
16	4.55	0.32	1.37	0.26	1.24	0.25	1.23											0.22	1.17
17	4.74	0.72	1.89	0.57	1.74	0.55	1.78											0.51	1.67
18	1.57	1.57	2.38	0.79	1.95	0.71	1.88											0.59	1.76
Σ			44.11		38.73		38.09												36.88

σ_s^2 = Variance of original (raw data) symbol values

σ_{-i}^2 = Theoretical variance utilizing prediction over i past frames

H_{-i} = Estimate of channel source entropy based upon expected minimum variance

SECTION IV

COMPUTER SIMULATION PROCEDURES AND PROCESSING SYSTEM MODELS

A. DESCRIPTION OF SPEECH DATA SAMPLES

Speech data samples for ten male speakers from the AFCRL Speech Library were used in this investigation. These samples were originally available in the form of punched paper tapes. Data in this form required special processing in preparation for utilization with digital computer programs. The inconvenience and general lack of reliability of this formatting was eliminated midway through the study when the data became available on magnetic tape. The magnetic tapes supplied by AFCRL contained data in six-character-per-word BCD format which is typical for use with IBM computers having internal binary logic. A program was written to convert these tapes to the five-character-per-word BCD format necessary for use with internal decimal logic machines. This conversion process is extremely fast and reliable, requiring no special purpose equipment.

There were 17 speech samples available for the ten speakers. These samples include text A for each speaker, and text B of seven of the ten. Texts A and B are shown in Figure 9. Text A contains 118 words and required about 35 seconds for recitation. Text B contains 108 words and required slightly less time for the average recitation. Table III provides the label descriptions for each sample available to the investigation. These labels were used to identify samples during simulation and evaluation.

The two texts made it possible to evaluate the ergodic-type assumption that source statistics accumulated from one message sample of sufficient length would be representative of statistics for any other message from the same source ensemble (speaker, in this case). Thus, predictors based on the statistics of one text were used to process the other, for each of the speakers on tape one.

The same speakers used in this study were used as subjects for the second study under this contract, a study of speaker recognition characteristics.[8] Comparisons were thus possible for processing one sample with predictors based on another sample, for various subjective estimates of similarity (or lack of similarity) between the speakers.

The entire texts (approximately 35 seconds) were processed for the sample covariance matrix and predictive coding simulations, for each speaker sample utilized. In most instances, a shorter speech sample would have sufficed for computation of the covariance matrix with about the same confidence in the estimates.

TEXT A

"Here is the sixth selection. It is a college lecture of an aspect of language.

We tend to think of a language as an accurate, stable thing, which we can use as we might a screwdriver or a pencil. It has a function and it will always serve that function well. Actually, even at a very low level, language can become slippery. We are not always sure what we will get when we order a Chef's Salad in a restaurant. When I ask for a Mexican Sundae in East Lansing, I get a "What's that" look; but I've discovered that if I ask for a Tin Roof, I get an object which is indistinguishable from a Mexican Sundae."

TEXT B

"A rose by any other name is still a rose; but one does have to know what a rose looks like. If I go to a nursery man to order a firebush, he probably should ask me some questions or at least take me into his grounds and point, saying, "Is that what you want?" "Or that?" If he doesn't, I'm apt to come home with an Acantha lalandi instead of a Folius alatus — hardly the same thing!

What I wish to do today is illustrate the semantic changes which occur in language — to make you more aware of the ambiguities which can arise when we use words."

Figure 9. Speech Sample Texts

B. CORE OF PREDICTIVE CODING SIMULATION PROGRAM

The block diagram of a typical predictive coding processing system was discussed in Section II. Figure 4 is a good reference for the following comments. The basic digital computer simulation program implements the processing scheme of that diagram. The output format for the results of the simulation will be described in this section. This format is independent of the model and parameters used for any one simulation. The Huffman Coder function and a description of its operating parameters is discussed in Section IV, paragraph C. The various options concerning parameter specification for the simulation

Table III. Sample Data Speaker Designations

Order of Sample	AFCRL Run No.	Talker	Text
		TAPE No. 1	
1	2	V0003	P06A
2	6	V0002	P06A
3	4	T0101	P06A
4	9	T0104	P06A
5	12	V0030	P06A
6	5	V0002	P06B
7	7	T0101	P06B
8	8	T0104	P06B
		TAPE No. 2	
1	13	V0048	P06A
2	15	V0038	P06A
3	16	V0037	P06A
4	19	V0019	P06A
5	23	V0046	P06A
6	26	V0014	P06A
7	14	V0048	P06B
8	17	V0037	P06B
9	18	V0038	P06B

program are described in Section IV, paragraph D. Four basic models were simulated for evaluation of the theory and effectiveness of predictive coding techniques included in this study. These models are described in Section IV, paragraph E. Finally, the simulation for analysis of buffer and delay requirements for Huffman Coding is described in Section IV, paragraph F. All of the features described in these sections are further details of the core simulation described in this section.

The output format for simulation is illustrated by the representative example, Tables IV and V. Additional labeling procedures were programmed for further identification, data, and program evolution referencing during the study. Simulation using any one of the speech samples required approximately 8 ±2 minutes computer time (IBM 7074).

The first output page of the simulation program, Table IV, provides information related to the apparent entropy of the transformed message. For each quantization level used in the simulation (number of levels arbitrarily chosen), the probability of occurrence of symbols from that level is given by the entry in the distribution table, referenced by the particular channel in question. In the same vertical alignment, the bottom line gives the frequency values for the sample as a whole; that is, independent of channel distinction. The former values guide the choice of coders if provided separately for each channel, while the later values would be the guide for a single coder used to process the entire symbol sequence.

The column labeled H(DIS) gives the average (channel) source entropy rate, computed directly from the frequency probabilities accumulated during simulation. H(NOR) is an estimate of the same quantity. It is based upon an approximation using probabilities obtained by quantization of values taken from a normal distribution having the simulation variance of the prediction errors. This estimate is of the same form as that computed along with the theoretical determination of optimum predictors, and serves for the comparisons. These comparisons substantiate the theory based on the assumption of normally distributed prediction errors, except when the variance is roughly less than one, in which case the errors deviate from the assumption of normality. In these highly desirable instances, H(NOR) is a much too conservative estimate.

The sums of the channel entropy values are printed below the respective column values. In the same column alignment, the H(DIS) and H(NOR) values are computed for the combined channel frequency distribution. These values are for comparison with the average channel entropy determined by dividing the sum by 18 (or, the combined channel values may be multiplied by 18 for comparison with the sum of separate channel values on a frame basis). The conclusion that a single Huffman Coder is adequate and efficient for predictive processing is based on comparisons of these average channel and combined channel entropy values.

The means are self explanatory. The expected values are zero, and any substantial deviation from zero is attributed to severe bimodel skewness of the original signal distribution. The simulation variance is also self explanatory. The theoretical variance is really of little interest to the problem at hand, but is of theoretical interest relative to the limiting case of quantization of continuous distributions. It is not very meaningful with the quantization widths used in the applications reported here. In some instances, the values are completely unrelated to the speech sample. The combined channel means and variances are computed to augment the combined channelfrequency distribution data.

Table IV. Computer Simulation Output Format

FREQUENCY DISTRIBUTION OF PREDICTION ERROR

RUN NUMBER 4 SPEAKER T0101 TEXT P06AS

NUMBER OF FRAMES USED IN SIMULATION 1354

CH. NO.	H(DIS)	H(NOR)	MEAN	SIMULATION VARIANCE	THEORETICAL VARIANCE	NORMALIZED FREQUENCY DISTRIBUTION								
1	2.50	2.55	0.001	2.0325	2.0885	0.028	0.018	0.065	0.227	0.301	0.234	0.097	0.023	0.007
2	2.07	2.09	-0.001	1.0643	0.9970	0.001	0.010	0.052	0.216	0.440	0.218	0.052	0.008	0.002
3	2.32	2.35	-0.004	1.5214	1.4690	0.003	0.020	0.064	0.243	0.362	0.215	0.065	0.018	0.010
4	1.97	2.05	0.004	0.9173	0.8678	0.000	0.001	0.042	0.243	0.440	0.223	0.041	0.007	0.003
5	2.05	2.08	-0.005	1.0451	1.0338	0.001	0.006	0.059	0.208	0.455	0.212	0.048	0.008	0.003
6	1.99	2.06	0.007	1.0170	1.0293	0.002	0.007	0.045	0.198	0.509	0.168	0.057	0.012	0.002
7	2.18	2.22	-0.004	1.2629	1.2173	0.003	0.015	0.052	0.225	0.428	0.199	0.061	0.013	0.006
8	2.39	2.40	0.000	1.6263	1.5117	0.003	0.022	0.083	0.223	0.342	0.222	0.078	0.022	0.006
9	2.40	2.41	0.006	1.6573	1.5714	0.003	0.015	0.085	0.253	0.324	0.188	0.107	0.021	0.005
10	2.36	2.37	-0.001	1.5775	1.4325	0.004	0.022	0.065	0.250	0.340	0.210	0.080	0.026	0.003
11	2.37	2.38	0.012	1.5879	1.5345	0.004	0.012	0.075	0.270	0.310	0.211	0.092	0.023	0.004
12	2.10	2.11	0.004	1.0871	0.9981	0.001	0.008	0.052	0.237	0.405	0.230	0.058	0.007	0.002
13	2.06	2.07	0.001	1.0258	0.9549	0.000	0.007	0.051	0.235	0.421	0.219	0.061	0.007	0.000
14	2.00	2.09	0.005	0.9727	0.9640	0.000	0.002	0.041	0.250	0.439	0.208	0.044	0.013	0.002
15	1.96	2.07	-0.006	1.0384	1.0132	0.003	0.018	0.032	0.190	0.515	0.198	0.028	0.010	0.007
16	1.09	1.43	0.016	0.3456	0.3680	0.000	0.000	0.009	0.097	0.795	0.074	0.018	0.006	0.001
17	1.01	1.34	-0.021	0.2939	0.3250	0.000	0.003	0.008	0.095	0.814	0.064	0.014	0.002	0.000
18	1.10	1.44	0.014	0.3508	0.3834	0.001	0.003	0.011	0.084	0.793	0.085	0.021	0.001	0.000
	35.92	37.50												
COMBINED CHANNELS	2.09	2.14	0.002	1.1347	1.0977	0.003	0.011	0.050	0.208	0.469	0.188	0.057	0.013	0.004

Table V. Computer Simulation Output Format

FREQUENCY DISTRIBUTION OF QUANTIZATION ERROR

RUN NUMBER 4 SPEAKER T0101 TEXT P06AS

NUMBER OF FRAMES USED IN SIMULATION 1354

CH. NO.	CRITICAL EPSILON	NO. GREATER THAN CRITICAL EPSILON	MEAN	VARIANCE ABOUT ZERO
1	0.500	20	0.003	0.1018
2	0.500	1	0.004	0.0806
3	0.500	4	0.002	0.0849
4	0.500	1	0.007	0.0857
5	0.500	1	-0.007	0.0841
6	0.500	2	0.008	0.0795
7	0.500	2	0.001	0.0873
8	0.500	1	-0.004	0.0833
9	0.500	2	0.003	0.0823
10	0.500	2	-0.006	0.0836
11	0.500	1	0.006	0.0849
12	0.500	0	0.008	0.0869
13	0.500	0	0.014	0.0831
14	0.500	0	0.011	0.0839
15	0.500	0	-0.010	0.0853
16	0.500	0	0.013	0.0727
17	0.500	0	-0.005	0.0723
18	0.500	0	0.009	0.0679

The second page of simulation output provides information related to the quantization noise present in the prediction process. The critical epsilon and number greater than critical epsilon values for each channel are for a comparison of the original and reconstructed symbol values at the receiver. For most, if not all, results included in this report, the width of quantization levels are normalized to unity, in agreement with the representation of speech sample data. Thus, referring to the ϵ_i notation of the system diagram, a ϵ_i value exceeding 0.5 in magnitude would correspond to a discrepancy between the original and the reconstructed symbol value, or spectrum amplitude value. Obviously, with unit quantization widths, the discrepancies are a consequence of the "clipping" that occurs at the extreme quantization levels. The percentage of such occurrences may be compared with the percentage of greatest magnitude values in the original message, quantized, but not subjected to a predictive coding transformation. This comparison is part of the overall system evaluation.

The mean and variance of the quantization noise are also computed for processing of each channel sequence. These statistics, shown in Table V, are provided for the interested reader who might, out of conditioned response, be urgently in need of a signal to noise ratio or dynamic range for the processing system. All sorts of ratios of averages and averages of ratios may be computed with the signal (prediction error) and noise (quantization error) variances. Similar gyrations may be performed with the number and half-width of the quantization levels.

C. VARIABLE LENGTH CODES — HUFFMAN CODER

At several points it has been implied that the Huffman Coder was a significant and integral part of the predictive coding scheme. The description of this section is to justify that implication.

The development in this report leans heavily on the utilization of entropy as a measure. The theory and evaluation of predictive coding are centered around the efficient utilization of a standard channel for which the binary symbol rate is the single measure of capacity. It has been explained that the probability distribution for symbols in an arbitrary alphabet is a form of measuring the uncertainty associated with processing messages composed of those symbols. Entropy was defined in terms of the probability distribution so as to represent the average number of binary symbols per message symbol required to "optimumly" transmit the message through the standard channel. This implication of optimum refers to a variable length coding scheme derived from the same probability distribution used to measure the entropy of the message symbols. Several variable length coding schemes have been evaluated to see how well they approach the minimum average bit rate measured by entropy. The various schemes, and estimates or formulas for computing their efficiency, have been reported in the literature of information theory.

Huffman Coding[9] is a method of assigning variable length codes to theoretically yield the lowest possible average message length (best approximation to the entropy measure), based upon the probabilities of symbols to be coded individually in sequence. Note that this sense of optimality excludes "run-length" coding, where long runs of a symbol value are coded to give symbol and length of run values. If the probability of any one symbol in the message alphabet greatly exceeds (say roughly, 0.70 probable) any other, run-length or "block" coding is necessary to efficiently transmit messages of those symbols. Block coding merely implies that one must group symbols enough to establish a set of probabilities for the groups that yeild more efficient codes for the group taken as an entity. Of course, the apparent entropy is decreased (at least not increased) when measured over the group probabilities. These concepts are explained in detail in the literature, and need not be pursued further for the present application. For typical results of predictive coding, the apparent entropy and code assignments are related to single symbol statistics, yielding codes 95-99 percent efficient.

Several examples of assigning Huffman Codes are given in the original publication explaining the method.[9] The assignment scheme is described by the following four steps:

1. Arrange the symbol probabilities in decreasing order.
2. Add the probabilities of the two lowest ranked symbols, thus forming a probability for the joint condition. Assign a zero code bit to one of the symbols, and a one code bit to the other. The order of assigning (upper or lower) the separate bits is arbitrary, but must be maintained as the procedure is repeated.
3. Rearrange the remaining and new joint-condition probabilities, again in decreasing order.
4. Repeat steps 2 and 3 until only the joint condition of all states remains. This state should have probability one, except for round-off error. Note that at each stage of the process, a "zero" or "one" code bit is assigned the symbol or group being combined. The variable length code for each symbol is determined by retracing the combination path of the probability measure for that symbol, prefixing at each stage the appropriate (0 or 1) code. This procedure is illustrated in Figure 10.

The entropy for the symbol probabilities used in Figure 10 and the average number of bits per message symbol generated by the Huffman Code are computed in Table VI. The entropy, 2.073, divided by the average binary code length, 2.107, gives an efficiency of 98.3 percent for this example code assignment. These results are typical for distributions with no single probability greatly exceeding one-half.

For a variable length code to be efficient at all, it must satisfy the prefix property. This property enables a coded message to be decoded without ambiguity. It means that the first few bits (prefix) of any code are not identical to a complete code word of the shorter length. There are other observations[10, 11] that are made concerning properties of variable length codes, but they need not be considered here. It should be apparent that any device for assigning variable

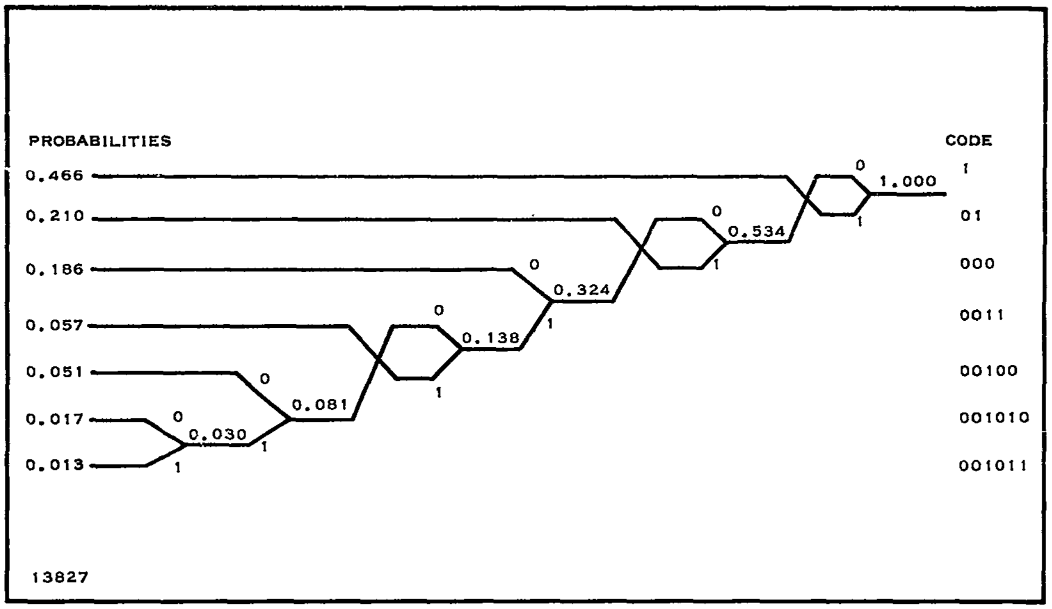


Figure 10. A Huffman Code Assignment

Table VI. Example Computations for Huffman Code Efficiency

l_i	p_i	$-p_i \log p_i$	$p_i l_i$
1	.466	.51334	.466
2	.210	.472823	.420
3	.186	.451352	.558
4	.057	.235574	.228
5	.051	.218961	.255
6	.017	.099931	.102
6	.013	.081449	.078
	1.000	2.073434	2.107

length codes requires an active memory (buffer) for storing message symbols or their codes. This buffer is necessary to smooth the talkspurts of variable lengths to the average mapping time and code length. A certain delay in processing of message symbols is inherent with the smoothing process. The buffer and delay requirements of a practical device depend upon so called high-order statistics of the message symbol series. That is, the characteristics associated with run lengths and cyclic orders within the series. Statistics of this type are difficult to accumulate. Analysis of the characteristics they measure is much more easily facilitated by computer simulation with the data. A computer simulation program to assist the analysis of buffer and delay requirements is described in Section IV, paragraph F.

The basic variable length coding ideas presented in this section are necessary for an understanding of the differences in processing models to be described in Section IV, paragraph E. To repeat, efficient variable length coding spells the difference between paper profit and the real world of practical efficient digital communication of speech messages.

D. PARAMETER OPTIONS FOR PREDICTIVE CODING SIMULATION PROGRAM

The options and reasons for options are described in roughly the order of occurrence in the simulated predictive coding process. An understanding of the options is necessary because of their relation to the basic prediction processing schemes and their effect on the measurements used for evaluation. Further, some of the options were utilized to perform simulations not included in this report. This latter point is especially true for the quantizer design parameters.

As the frame data samples are introduced into a simulation sequence, one of two operations can be performed. In most models, the channel mean amplitude values were subtracted from the respective channel symbol values for that frame. This operation accomplishes the result mentioned in Section III, paragraph A. The expected mean value for all symbols of the resulting input and transformed output sequences will be zero. This consistance is a matter of convenience to be preferred over keeping track of variations from expectation for 18 different channel average values. It provides a common alphabet for values in each vocoder channel. When processing one speech sample with prediction based on statistics of another, any significant difference in the statistics will result in a shift of the channel means away from the expected zero value, thus, usually increasing the entropy of the combined channel series.

The second operation available at this point is a frame by frame subtraction. That is, the symbol values of the respective channels from the previous frame are subtracted from the values of the current frame. Processing schemes using this operation are referred to as "differenced data." The expected values of the differenced symbol series are also zero before and after predictive coding.

Subtraction of means and differencing are linear operations (of prediction) in themselves, but are also followed by the predictive coding transformations to eliminate linear intersymbol influence extending across channels and further back in time. The simpler operations just described do not lower the apparent entropy to the extent of predictive coding, if fidelity is preserved.

When simulating with unequal length predictors, the order of multiplexing effects the results. An option for multiplexing in two orders is available in the basic program.

Once the data has been multiplexed, a logical decision switch is tested to accept data on silence-nonsilence (VAP) and voiced/unvoiced (V/UV) options. The basic processing schemes discussed in the next section rely on these choices. When VAP = 0, all channel amplitude values are zero for that frame. The V/UV parameter denotes the character of speech energy in nonsilence frames. Since the source statistics are different for each classification of signal structure, prediction schemes based on these differences have been investigated.

The parameters associated with quantization of the error (transformed) signal have direct effects on the entropy and noise measures. On a phychacoustic basis, the preservation of source fidelity would also be affected by coarse quantization. Both the range and widths of quantization levels may be arbitrarily chosen with one constraint. The simulation output format of symbol (level) distribution statistics allows no more than nine levels to be printed on a single line with the other channel data. No provision for multiple lines per channel was deemed necessary, consequentially, not programmed.

Although simulation with several speech samples was performed with many variations in number and width of quantization levels, only the widths identical to those used in recording the raw data are included in this report. Without facilities to resynthesize the transformed signal, no means of comparing the source fidelity was available for data processed with arbitrary quantization widths.

The analysis of buffer and delay characteristics of Huffman Coding requires transformed message data on magnetic tape for use with a separate computer program. An option was provided for writing this tape during the processing simulation. The simulation time is increased slightly by the preparation of these tapes. This transformed message data was further analyzed for linear independence of message symbols.

The critical epsilon values discussed in Section IV, paragraph B are part of the optional data. They are equal and constant for each channel in the simulation results included in this report. In the event a basis for comparing simulation results and a subjective fidelity measure is developed in the future, this option would be helpful. These values were varied with variations in quantization widths.

When prediction is based on a V/UV option, so must the logic for addressing predictor coefficient arrays and reinitialization of the active memory containing past symbol values. These operations are all based on a single logic. Of course, initial data input must be assembled in accordance with this processing model logic.

E. FOUR BASIC PROCESSING SCHEMES

Four basic processing models were used for the evaluation of predictive coding techniques investigated during this study. Undoubtably there are other schemes of comparable complexity that were not envisioned in the attempt to cover so much ground. These models do demonstrate trade-offs in utilization of recognized source statistical structure. Recall that all processing operations effect only the spectrum data of the vocoded speech source.

1. Model 1

In several respects, the simplest model is to process every source symbol of every frame independently of voicing of silence characteristics. For this model, all source sample data for a message is used to compute the sample covariance matrix, Σ. The ELP or ULP are computed from Σ and used for prediction. The greatest reductions in the signal variance are realized for predictive transformations on this type of source data. Reductions in the variance by factors from 3 to 15 are common, depending on the channel and length of predictors. Roughly, the results are weighted considerably by the effect of silence predicting silence. A model of this type was the first to be investigated in the study.

Huffman Coding of message processed in this manner requires less buffer and delay than for the other schemes. The anticipated performance differences between design and implementation with arbitrary messages are in a favorable direction. The tendency of silence to predict silence efficiently should result in a message probability exceeding design probability for the most likely symbol, namely, zero prediction error. This result would reduce the tendency of buffer overflow and similarly reduce the actual information transmission rate, allowing for more synchronization codes. Any silence would tend to stabilize the prediction accuracy. Conversely, channel noise could have more of an effect on fidelity of transmission. The pros and cons of this type have not been studied in detail.

Predictive coding in this manner has been demonstrated consistantly to reduce the average bit rate to less than two-thirds of the original rate. This rate is certainly weighted by the percentage of silence in the source message, but not in a manner that is simple to estimate. Furthermore, in a practical hardware application, greater than anticipated reduction in average message bit rate due to increased silence is not easily accomplished with this model. The importance of this consideration in evidenced in Model 2.

2. Model 2

This model is distinguished from the previous one in the recognition and processing of silence-nonsilence information. The voice amplitude parameter (VAP) of each vocoded data frame serves as the decision variable. When VAP is zero, all spectrum amplitude values are zero for that frame. Transmission of the VAP value unquestionably accounts (in the most efficient manner) for the major information in a total message. Remember, the information theoretic measure is used and not a subjective ordering of what speech structure is more major than minor, etc.

The principal concept of this model is that of spreading speech burst (nonsilence) spectrum information over the silence intervals. The "information" regarding silence or nonsilence is transmitted independent of the spectrum data by the VAP value. The hardware implications principly effect control logic, rather than processing time or component expense.

The hypothetical graph shown in Figure 11 illustrates the central idea for this model. The average source information bit rate for predictive coded speech burst data is represented by the slope of the thick oblique line segments. The message average information bit rate is represented by the slope of the dashed line. The horizontal segments during silence intervals indicate no contribution to the spectrum information being (transmitted) generated.

As a practical consideration, this processing model is very adaptive to efficient utilization of limited channel capacity. Contrary to the limitation of Model 1, the influence of silence in a message is explicitly part of the control on average bit rate. A warning light, buzzer, or distractive attention-getter of some type on the vocoder might easily be activated by a sensing connection to the Huffman Coder buffer. Any pause by the speaker is applied optimumly toward the alieviation of any tendency to exceed the preset average channel bit rate.

In a sense, Model 1 has implicitly this, "control by silence." However, other factors of that model, such as the effect of increased silence to shift the mean of symbol values, make it very difficult to estimate the dependence of bit rate on the single parameter, percent silence. Further, a study of Model 2 provides measures and insight for the speech structure that is least subject to efficient processing by methods other than predictive coding. Inginuity may be exercised for ways of applying the results presented here concerning speech burst information bit rates.

Figure 11 is also helpful for describing buffer and delay requirements of the Huffman Coder for this scheme. The type of simulation described in Section IV, paragraph F provides operational data similar to that characterized by the graph. Graphs of this type could also be formed with average statistics to typify any reasonably defined source ensemble. Recognizing that there are engineering details that distinguish between hardware and the graph

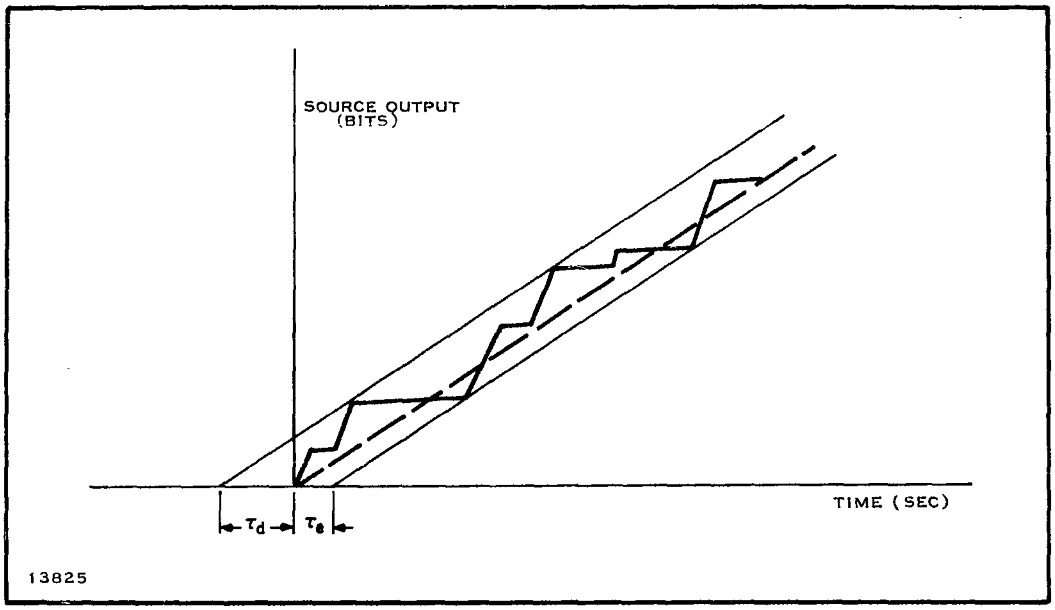


Figure 11. Graphical Representation of Model 2

characterization of performance, the graph still provides the most significant type of information, and in a form to provide good estimates for buffer and delay implementation.

The light lines drawn parallel to the message average rate may be used to describe the transmitter-receiver, buffer, and delay interplay. The lower line may be associated with the actual transmission of data. The upper line (same rate) may similarly be associated with receiving and decoding of data. Note that a combined buffer capacity (in bits) equal to the vertical separation of the two lines is "adequate." The horizontal intercepts are related to delay in the system necessitated by the variable length codes and the smoothing of speech burst data over the silence intervals. The intercept τ_c is an estimate of the delay at the coder, and τ_d is an estimate of delay at the decoder. As the percentage of silence used to determine the transmission rate increases, so typically do the buffer and delay requirements.

The simulations related to this model utilize the statistics and data for speech bursts only. The message average bit rate is thus estimated by multiplying the bit rate during speech bursts by the percentage of speech bursts time for the message to be processed. In applications where long delays are acceptable (any one-way channel for instance), the speech burst information may be spread out over much silence to achieve extremely low message bit rates.

3. Model 3

Both schemes described so far can be used with the processes of Models 3 and 4. That is, the models are not mutually exclusive.

The processing with differenced data is labeled Model 3. As previously mentioned, differencing is a form of nonstatistical linear prediction. It theoretically doubles the necessary alphabet. It is a restricted form of linear prediction that does not have the entropy reducing expectation of the linear statistical prediction using more data from the message past. However, differencing followed by predictive coding utilizes past symbol influence in a manner different from the other models. For this reason, it was considered in the study for comparison.

4. Model 4

Model 4 is a scheme for utilizing the voiced and unvoiced structure of speech burst data. Because the energy of voiced speech is predominantly in the lower spectrum channels, and voiced speech energy is not, the statistics of the source vary with this classification of structure.

Separate predictors are used for the voiced and unvoiced structure, instead of the single predictor for all speech burst data for a channel. Also, the respective means are subtracted from the input data, rather than the means for all speech burst data. The logic and complexity of the processing system are increased substantially by this scheme. The accuracy of prediction increases with this scheme, but not in proportion to the increased complexity over the simpler approaches. This is due primarily to the small precentage of unvoiced structure, and will be discussed in the analysis of results.

F. HUFFMAN CODING SIMULATION FOR BUFFER AND DELAY ANALYSIS

In this section, the computer simulation program utilized for the predictive coding source rate analysis is described. Essentially, the program is used to provide simulation data of the type illustrated by Figure 11. The interpretation of simulation results is described in Section IV, paragraph E under the discussion of the Model 2 predictive coding scheme. The program provides data related to the size of buffer requirements, the coding and decoding delays of message processing, and measures of the average coding rate and efficiency.

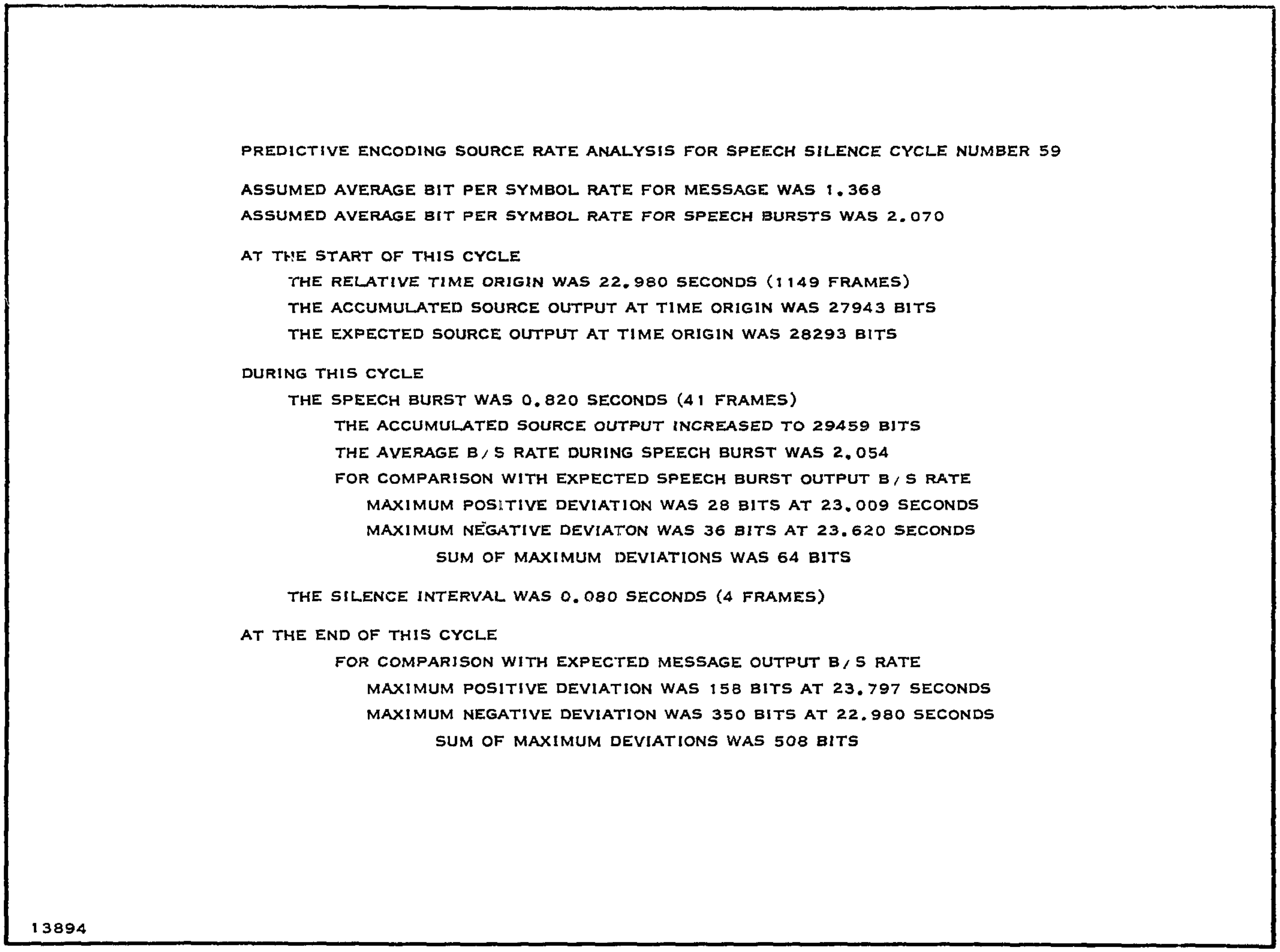
As an option in the basic predictive coding simulation program, the transformed message may be written in frame groups on magnetic tape for use in this simulation program. At the time this tape is written, the silence-nonsilence information is also recorded. For convenience in the output of simulation results, and for adaptibility to the Models 1 and 2 processing schemes, the logic is performed in message cycle groups. A cycle is a speech burst and its following silence interval. When processing according to Model 1, all coding silence intervals are of length zero. The entire message cycle is considered a coding speech burst. This logic results in a small amount of redundancy in the output for Model 1, but satisfies the analysis requirements for both models.

Figures 12 and 13 are typical simulation outputs for processing Model 2. They may serve as illustrations for an explanation of the output format. Figure 11 is representative of output at the end of each speech cycle. Figure 12 is an example of the summary output upon completion of the coding simulation for an entire message. Most of the labeling is self explanatory. Only a few observations are included here.

The cycle outputs were formated for convenience in plotting the results for any arbitrary length of the simulated message. The basic time unit, the sampling rate, of the vocoded source is given on the summary page. A program identification page (not shown here) was output at the beginning of each simulation. This page described the Huffman Code assignment scheme used in the simulation, and additional labeling to describe the speaker and predictive coding simulation identification. The code assignment scheme was developed by hand in the manner described in Section IV, paragraph C.

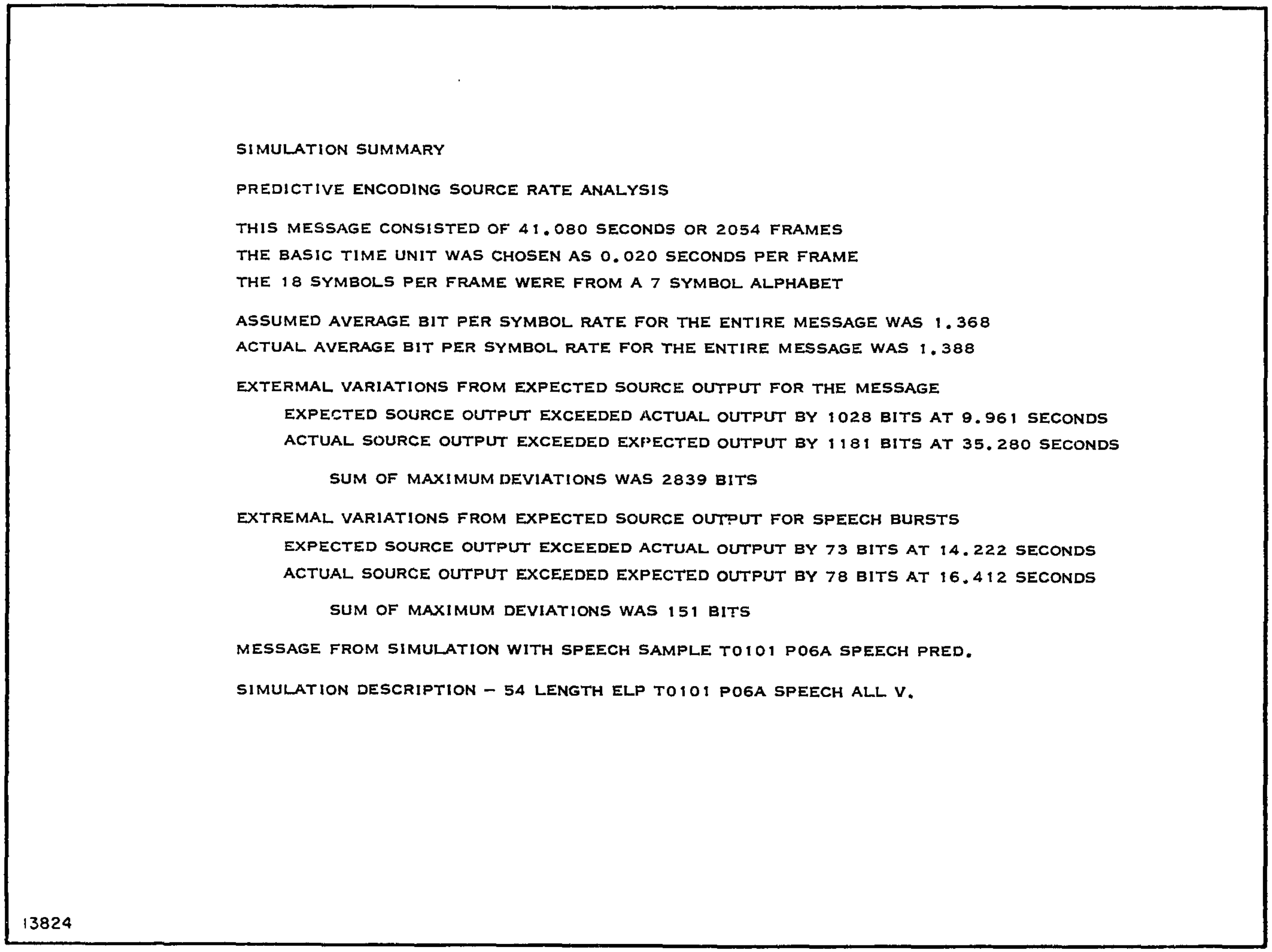
The assumed average bit per second rate during coding speech bursts is just the entropy of the combined channel statistics that was computed during the predictive coding simulation. For coding of Model 2 data, the assumed average bit per second rate for the message was computed by multiplying the former rate by the fraction of nonsilence time for the speech sample. In both cases, entropy measures are used, and no allowance is made for the less than perfect efficiency of the Huffman Code assignment. Note from the actual average rates of simulation given on the summary page, the message average coding efficiency is computed to be 1.368 divided by 1.388 = 0.986.

The sum of maximum absolute deviations between actual and assumed bit production was computed during each cycle for both the speech burst and message performance. This estimate of adequate buffer size was mentioned in Section IV, paragraph E. The unmentioned output is considered self explanatory.



```
PREDICTIVE ENCODING SOURCE RATE ANALYSIS FOR SPEECH SILENCE CYCLE NUMBER 59

ASSUMED AVERAGE BIT PER SYMBOL RATE FOR MESSAGE WAS 1.368
ASSUMED AVERAGE BIT PER SYMBOL RATE FOR SPEECH BURSTS WAS 2.070

AT THE START OF THIS CYCLE
   THE RELATIVE TIME ORIGIN WAS 22.980 SECONDS (1149 FRAMES)
   THE ACCUMULATED SOURCE OUTPUT AT TIME ORIGIN WAS 27943 BITS
   THE EXPECTED SOURCE OUTPUT AT TIME ORIGIN WAS 28293 BITS

DURING THIS CYCLE
   THE SPEECH BURST WAS 0.820 SECONDS (41 FRAMES)
      THE ACCUMULATED SOURCE OUTPUT INCREASED TO 29459 BITS
      THE AVERAGE B/S RATE DURING SPEECH BURST WAS 2.054
      FOR COMPARISON WITH EXPECTED SPEECH BURST OUTPUT B/S RATE
        MAXIMUM POSITIVE DEVIATION WAS 28 BITS AT 23.009 SECONDS
        MAXIMUM NEGATIVE DEVIATON WAS 36 BITS AT 23.620 SECONDS
             SUM OF MAXIMUM  DEVIATIONS WAS 64 BITS

   THE SILENCE INTERVAL WAS 0.080 SECONDS (4 FRAMES)

AT THE END OF THIS CYCLE
      FOR COMPARISON WITH EXPECTED MESSAGE OUTPUT B/S RATE
        MAXIMUM POSITIVE DEVIATION WAS 158 BITS AT 23.797 SECONDS
        MAXIMUM NEGATIVE DEVIATION WAS 350 BITS AT 22.980 SECONDS
             SUM OF MAXIMUM DEVIATIONS WAS 508 BITS
```

Figure 12. Huffman Coding Simulation-Typical Cycle Format



```
SIMULATION SUMMARY

PREDICTIVE ENCODING SOURCE RATE ANALYSIS

THIS MESSAGE CONSISTED OF 41.080 SECONDS OR 2054 FRAMES
THE BASIC TIME UNIT WAS CHOSEN AS 0.020 SECONDS PER FRAME
THE 18 SYMBOLS PER FRAME WERE FROM A 7 SYMBOL ALPHABET

ASSUMED AVERAGE BIT PER SYMBOL RATE FOR THE ENTIRE MESSAGE WAS 1.368
ACTUAL AVERAGE BIT PER SYMBOL RATE FOR THE ENTIRE MESSAGE WAS 1.388

EXTERMAL VARIATIONS FROM EXPECTED SOURCE OUTPUT FOR THE MESSAGE
    EXPECTED SOURCE OUTPUT EXCEEDED ACTUAL OUTPUT BY 1028 BITS AT 9.961 SECONDS
    ACTUAL SOURCE OUTPUT EXCEEDED EXPECTED OUTPUT BY 1181 BITS AT 35.280 SECONDS

        SUM OF MAXIMUM DEVIATIONS WAS 2839 BITS

EXTREMAL VARIATIONS FROM EXPECTED SOURCE OUTPUT FOR SPEECH BURSTS
    EXPECTED SOURCE OUTPUT EXCEEDED ACTUAL OUTPUT BY 73 BITS AT 14.222 SECONDS
    ACTUAL SOURCE OUTPUT EXCEEDED EXPECTED OUTPUT BY 78 BITS AT 16.412 SECONDS

        SUM OF MAXIMUM DEVIATIONS WAS 151 BITS

MESSAGE FROM SIMULATION WITH SPEECH SAMPLE T0101 P06A SPEECH PRED.

SIMULATION DESCRIPTION - 54 LENGTH ELP T0101 P06A SPEECH ALL V.
```

13824

Figure 13. Huffman Coding Simulation – Summary Format

SECTION V

ANALYSIS OF STUDY PROGRAM RESULTS

The following paragraphs present the major accomplishments, observations, and recommendations concerning the primary objectives of this study. The conclusions presented are certainly not exhaustive, but an effort has been made to tabulate sufficient data to enable the reader to make additional observations. Results of over fifty digital computer predictive coding simulations are presented.

A. TABULATION OF PREDICTIVE CODING SIMULATION DATA

Results of 35 computer simulations are compiled as Table VII. These results were extracted from the simulation output data for which the format was described in Section IV. This primary reference table and the further results in Table VIII, will serve for observations and conclusions pertaining to the majority of predictive coding models of interest. It is believed that sufficient data is presented to enable the reader to verify and augment the comments on the analysis.

Table VII contains representative statistics for 17 speech data/predictor data combinations as described in the left-hand column. Note that items 7 and 15 are identical, the repetition is for convenience in grouping relevant data.

The 21 statistics are identified by the column headings, most of which are self explanatory. The "Avg. Chan. H" descriptor refers to the average channel symbol entropy values for each channel. The "Comb. Chan H" descriptor refers to the symbol entropy computed from the combined channel simulation statistics. Models 1 and 2 are the "all data" and "speech burst data" processing schemes described in Section IV. The entropy values for the Model 2 scheme are computed as the product of the speech burst entropy and the fraction of speech burst data in the text sample. The compression factors given in columns 20 and 21 are ratios of the simulation entropy and the 3 bit/symbol value representing the H_{max} for eight-level quantization of the original data. For data represented in this table with seven-level quantization, the average amount of "clipping" (number exceeding critical epsilon) was in all cases less than 2 percent, and more typically near 1/2 percent.

A more complete coverage of the dependence upon length and number of quantization levels is presented in the data of Table VIII, for a single speaker. Note that results of simulation with differenced data (model 3) are included for comparison. The "message percent clipping values" were computed by summing the total "clips" over all channels and dividing by the total number of message symbols. The channel values were computed similarly, using measures associated with a single channel.

Table IX provides entropy measures as described by the column headings. The raw data was quantized with eight levels. The low entropy of the raw data for the top four channels results from voiced data having little or no energy in these channels. The estimated entropy values were computed along with computation of predictor coefficients. The quantized-normal approximation described in Section III was used for each expected variance associated with the predictive transformation. Combined channel values for the simulation entropy are included for comparison. The conservative bias of the theoretical estimates for small variances is illustrated particularly in the high channels.

Note how the effectiveness of prediction varies over the channels . The channels with poorest prediction effectiveness vary somewhat with speakers as do the formant structure characteristics to which it is attributed. A similar, but less pronounced, variation is apparent in the raw data. The higher entropy values result from a "flatter" amplitude probability distribution, conceivably a result of time-varying formant migration across the channel, or channels, in question. The combined channel symbol distributions for three types of data are illustrated in Figure 14. These distributions are provided with their entropy values as representative of those of similar classification. The raw data symbol levels are shown out of order in Figure 14a to correspond more closely with the distributions having zero mean. The distribution of prediction errors, Figure 14b, has the symmetry and weighting typical of normally distributed and quantized values. As has been observed with most differenced data, the distribution in Figure 14 c is very symmetric and has a large zero level probability but is not weighted with gaussian envelope. The entropy values are typical of each classification.

Figures 15 through 18 illustrate the distributions of symbol values over each vocoder channel. The shape (distribution) of the graphs for the raw data shown in Figures 15 through 17, will vary over speaker classifications. The prediction error distributions do not vary significantly in this manner. In any event, these examples are presented solely for illustration. From the first three, the difference in the voiced and unvoiced distribution characteristics are vividly shown. Based upon the speaker recognition study[8], speaker T0101 was judged on a subjective basis to have a generally "dull," "low," and "resting" voice. A speaker (such as V0019) with a generally "intense," "high," and "busy" voice has considerably different distributions across the channels.

Finally, for the graphical and tabulated illustrations from prediction simulations, Figure 19 is a plot of the message average entropy measures relative to the percent silence for most of the Model 2 simulation data. The graph is self explanatory, but will be disucssed in the following paragraph. The hybrid simulations are easy to pick out of Table VII as they result from simulation with one speaker using predictors derived from the statistical characteristics of another.

1

	Simulation Model and Statistics Text Simulation Sample Data							Model 1 54 ELP 7 level		Speech Burst 54 ELP 7 level		Speech Burst
		Length of sample (sec)	Percent silence	Mean silence interval (sec)	Stand dev-silence (sec)	Mean speech interval (sec)	Stand dev-speech (sec)	Avg. Chan H	Comb Chan H	Avg Chan H	Comb Chan H	Avg Chan H
		1	2	3	4	5	6	7	8	9	10	11
1.	V0003-A	35	15	0.09	0.14	0.49	0.89			1.99	2.07	
2.	V0030-A	35	21	0.09	0.18	0.35	0.47			2.01	2.10	
3.	V0002-A	34	27	0.10	0.17	0.26	0.33	1.90	1.99	2.04	2.14	2.06
3.*	V0002-A-Diff Data-No Prediction							1.81	1.88			
4.	V0002-A (ULP)	34	27	0.10	0.17	0.26	0.33			2.05	2.13	2.07
5.	V0002-B	31	26	0.12	0.20	0.35	0.45			2.00	2.10	
6.	V0002-B (V0002-A Pred)	31	26	0.12	0.20	0.35	0.45			2.05	2.16	
7.	T0101-A	41	34	0.13	0.24	0.26	0.33	1.77	1.85	1.96	2.07	2.00
8.	T0101-B	35	30	0.12	0.23	0.29	0.37			1.95	2.04	
9.	T0101-B (T0101-A Pred)	35	30	0.12	0.23	0.29	0.37	1.82	1.90	2.00	2.09	2.01
10.	T0104-A	32	23	0.08	0.13	0.25	0.35	1.88	1.98	1.96	2.05	
11.	T0104-B	29	23	0.08	0.14	0.28	0.37			1.93	2.03	
12.	T0104-B (T0104-A Pred)	29	23	0.08	0.14	0.28	0.37			2.00	2.10	
13.	V0048-A	42	27							2.00	2.07	
14.	V0048-B (V0048-A Pred)	42	30							2.04	2.12	
15.	T0101-A (Same as No. 7)	41	34	0.13	0.24	0.26	0.33	1.77	1.85	1.96	2.07	2.00
16.	T0101-A (V0019-A Pred)	41	34	0.13	0.24	0.26	0.33			2.12	2.28	
17.	V0019-A	47	33							2.06	2.15	2.08
18.	V0019-A (T0101-A Pred)	47	33							2.20	2.30	2.29

Table VII. Results of Computer Simulations

	Model 1 54 ELP 7 level		Speech Burst 54 ELP 7 level		Speech Burst 54 ELP 9 level		Speech Burst 18 ELP 7 level		Model 2 54 ELP 7 level		Model 2 18 ELP 7 level		54 ELP 7 level	Model 2 54 ELP 7 level	Model 2 54 ELP 7 level
Stand dev-speech (sec)	Avg. Chan H	Comb Chan H	Avg Chan H	Comb Chan H	Avg Chan H	Comb Chan H	Avg Chan H	Comb Chan H	Avg Chan H	Comb Chan H	Avg Chan H	Comb Chan H	H ratio = $\frac{\text{Mod 2 H}}{\text{Mod 1 H}}$	Comb Chan Compression Factor	Avg Chan Compression Factor
6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
.89			1.99	2.07			2.00	2.08	1.07	1.76	1.70	1.77		0.586	0.566
.47			2.01	2.10			2.06	2.16	1.59	1.66	1.63	1.71		0.553	0.530
.33	1.90	1.99	2.04	2.14	2.06	2.16	2.09	2.20	1.48	1.56	1.52	1.61	0.78	0.520	0.493
	1.81	1.88													
.33			2.05	2.13	2.07	2.15			1.50	1.55				0.516	0.500
.45			2.00	2.10					1.48	1.55				0.516	0.493
.45			2.05	2.16					1.52	1.60				0.533	0.507
.33	1.77	1.85	1.96	2.07	2.00	2.09	2.05	2.14	1.31	1.37	1.35	1.41	0.74	0.457	0.436
.37			1.95	2.04					1.37	1.43				0.477	0.456
.37	1.82	1.90	2.00	2.09	2.01	2.12			1.40	1.46			0.77	0.486	0.466
.35	1.88	1.98	1.96	2.05			2.01	2.12	1.51	1.58	1.55	1.63	0.80	0.526	0.504
.37			1.93	2.03					1.49	1.56				0.520	0.496
.37			2.00	2.10					1.54	1.61				0.536	0.514
			2.00	2.07					1.46	1.51				0.503	0.486
			2.04	2.12					1.43	1.48				0.494	0.476
.33	1.77	1.85	1.96	2.07	2.00	2.09	2.05	2.14	1.31	1.37	1.35	1.41	0.74	0.457	0.436
.33			2.12	2.28					1.40	1.50				0.500	0.467
			2.06	2.15	2.08	2.18			1.38	1.44				0.480	0.460
			2.20	2.30	2.29	2.42			1.47	1.54				0.514	0.490
														Avg. =	0.490
														(No. 15 excluded)	

2

Table VIII. Predictive Coding Simulation Statistics

Text Speech Data

Sample 4T0101 P06A

Speech Burst Data

length level	18	36	54
5	2.03	1.98	1.98
7	2.14	2.07	2.07
9	2.16	2.09	2.09

Differenced Speech Burst Data

length level	0	18	36
5	1.94	1.97	1.99
7	2.14	2.11	2.11
9	2.22	2.16	2.15

Combined Channel Apparent Entropy

length level	18	36	54
5	4.55	3.41	3.34
7	0.99	0.80	0.70
9	0.21	0.16	0.15

length level	0	18	36
5	8.61	5.31	4.82
7	4.48	1.68	1.42
9	2.25	0.52	0.40

Message Total Percent Clipping

length level	18	36	54
5	9.90	8.58	8.35
7	3.77	3.40	2.88
9	1.55	1.25	1.48

length level	0	18	36
5	12.91	9.66	9.37
7	8.56	5.17	4.87
9	5.02	2.36	2.14

Channel Maximum Percent Clipping

Table IX. Frame Entropy Comparisons

Raw Data, Theoretical Estimates, and Simulation Results

Seven-Level Quantization-Speech Burst Data-Sample 4T0101 P06A

		18 ELP		36 ELP		54 ELP	
Chan.	Raw Data	Estimate	Simulation	Estimate	Simulation	Estimate	Simulation
1	2.886	2.54	2.45	2.52	2.41	2.52	2.42
2	2.745	2.22	2.26	2.04	2.11	2.03	2.11
3	2.887	2.40	2.39	2.31	2.30	2.30	2.24
4	2.739	2.04	2.06	1.93	1.95	1.92	1.97
5	2.714	2.11	2.10	2.07	2.08	2.05	2.06
6	2.372	2.11	1.99	2.05	1.98	2.05	1.96
7	2.611	2.26	2.22	2.19	2.13	2.17	2.20
8	2.817	2.41	2.41	2.33	2.33	2.32	2.35
9	2.828	2.42	2.44	2.36	2.37	2.35	2.38
10	2.830	2.39	2.41	2.31	2.33	2.28	2.31
11	2.817	2.43	2.43	2.35	2.36	2.33	2.34
12	2.760	2.15	2.17	2.03	2.07	2.03	2.09
13	2.835	2.10	2.13	2.01	2.07	1.99	2.06
14	2.571	2.16	2.16	2.01	2.01	2.00	2.02
15	1.080	2.09	1.99	2.06	1.94	2.04	1.93
16	0.962	1.30	1.07	1.25	1.08	1.24	1.09
17	0.969	1.19	1.05	1.15	1.01	1.14	1.00
18	0.463	1.42	1.10	1.30	1.03	1.28	1.08
	41.879	37.77	36.83	36.27	35.57	36.04	35.66
Combined Channel							
	46.879		38.52		37.26		37.26

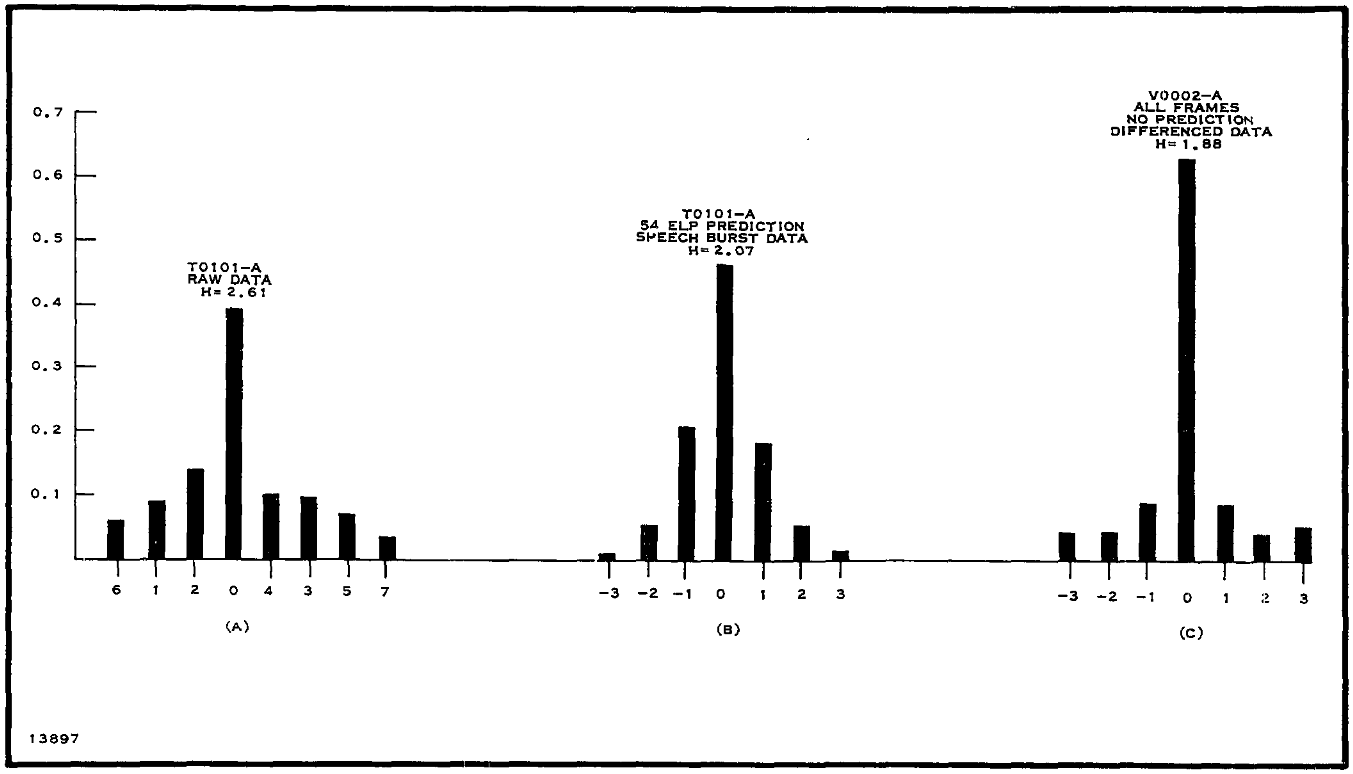


Figure 14. Combined Channel Symbol Distributions

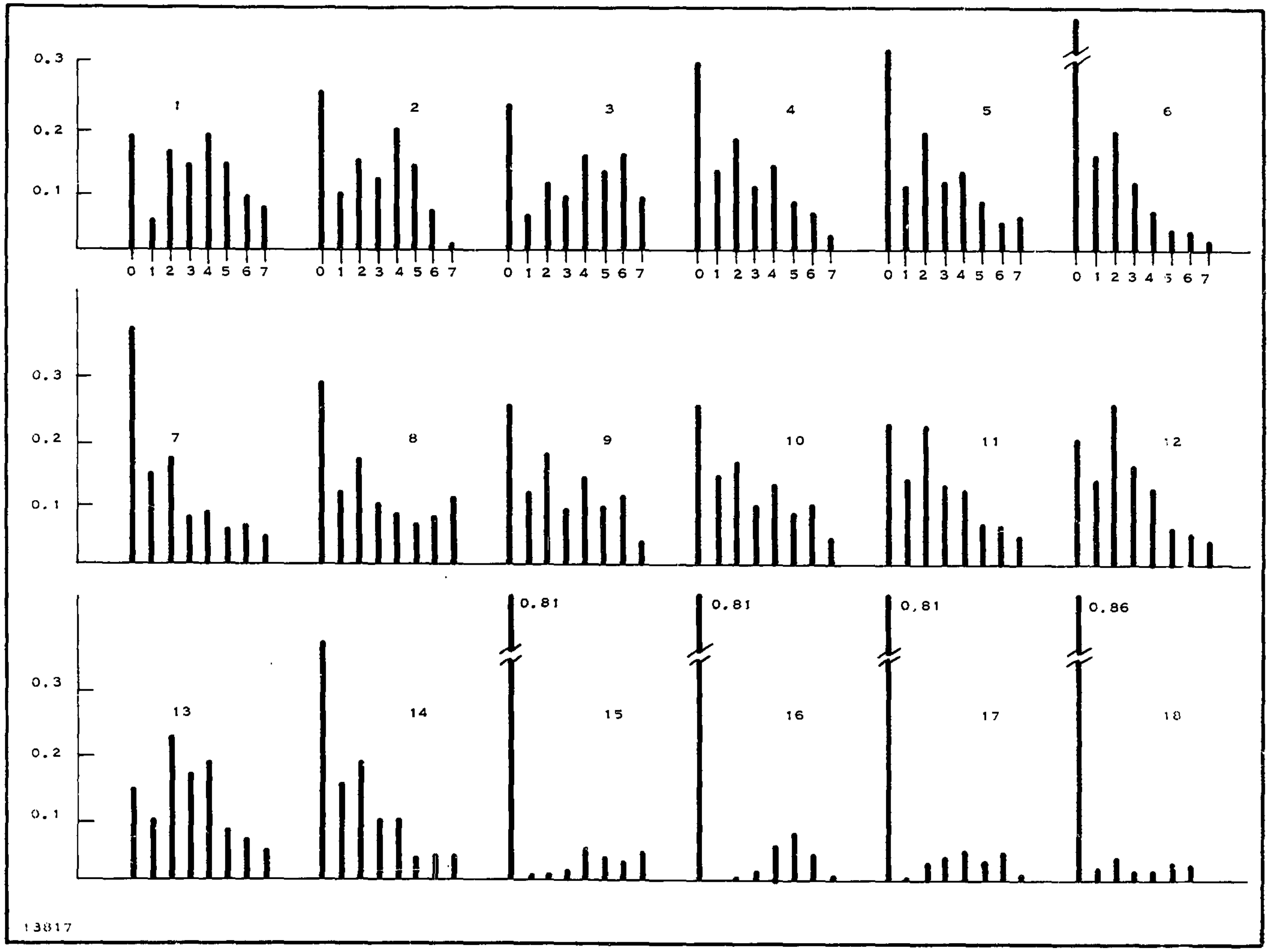


Figure 15. Distribution of Channel Values-0004 T0101 P06AS-Nonsilence Frames

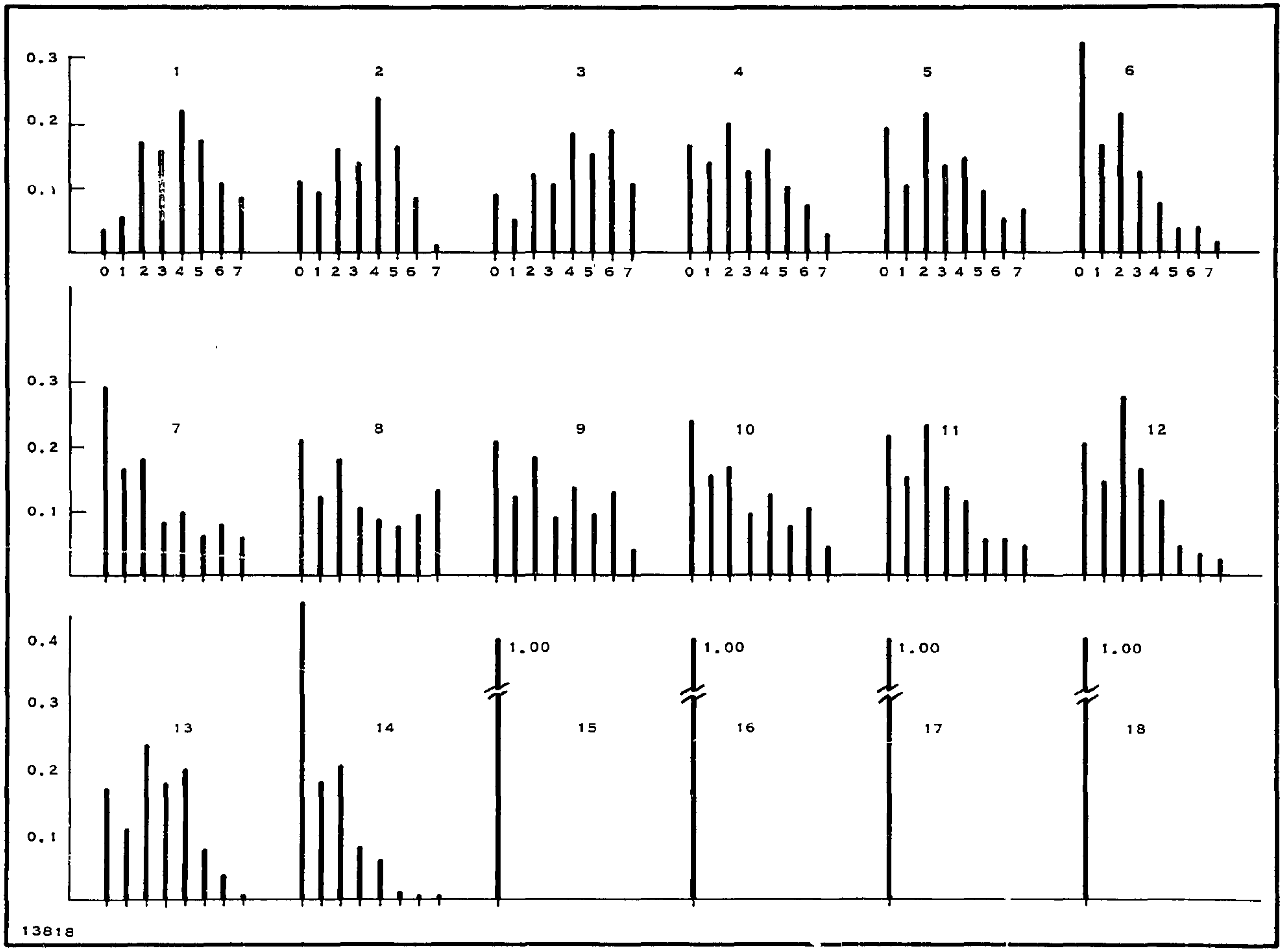


Figure 16. Distribution of Channel Values-0004 T0101 P06AS-Voiced Frames

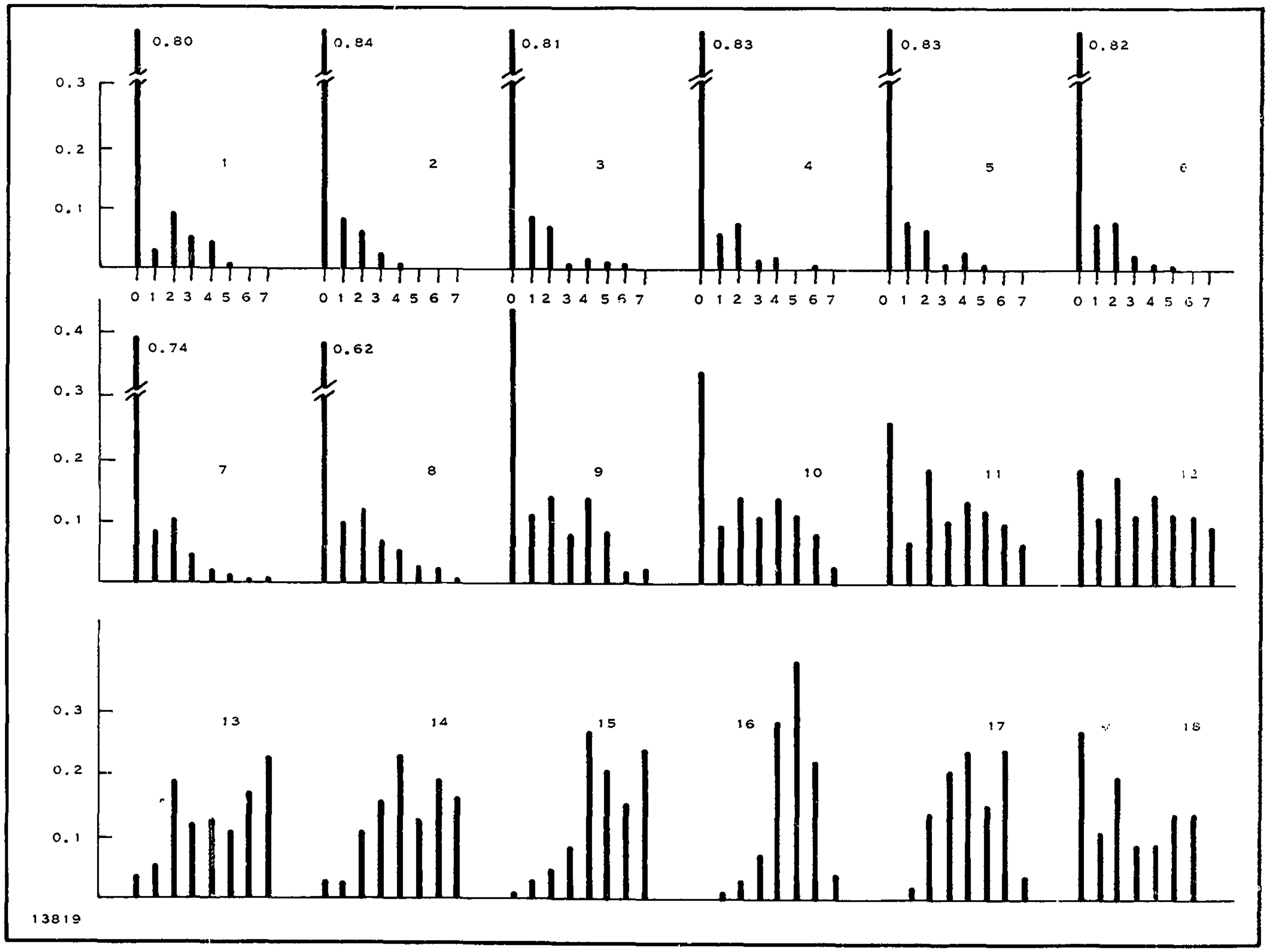


Figure 17. Distribution of Channel Values-0004 T0101 P06AS-Unvoiced Frames

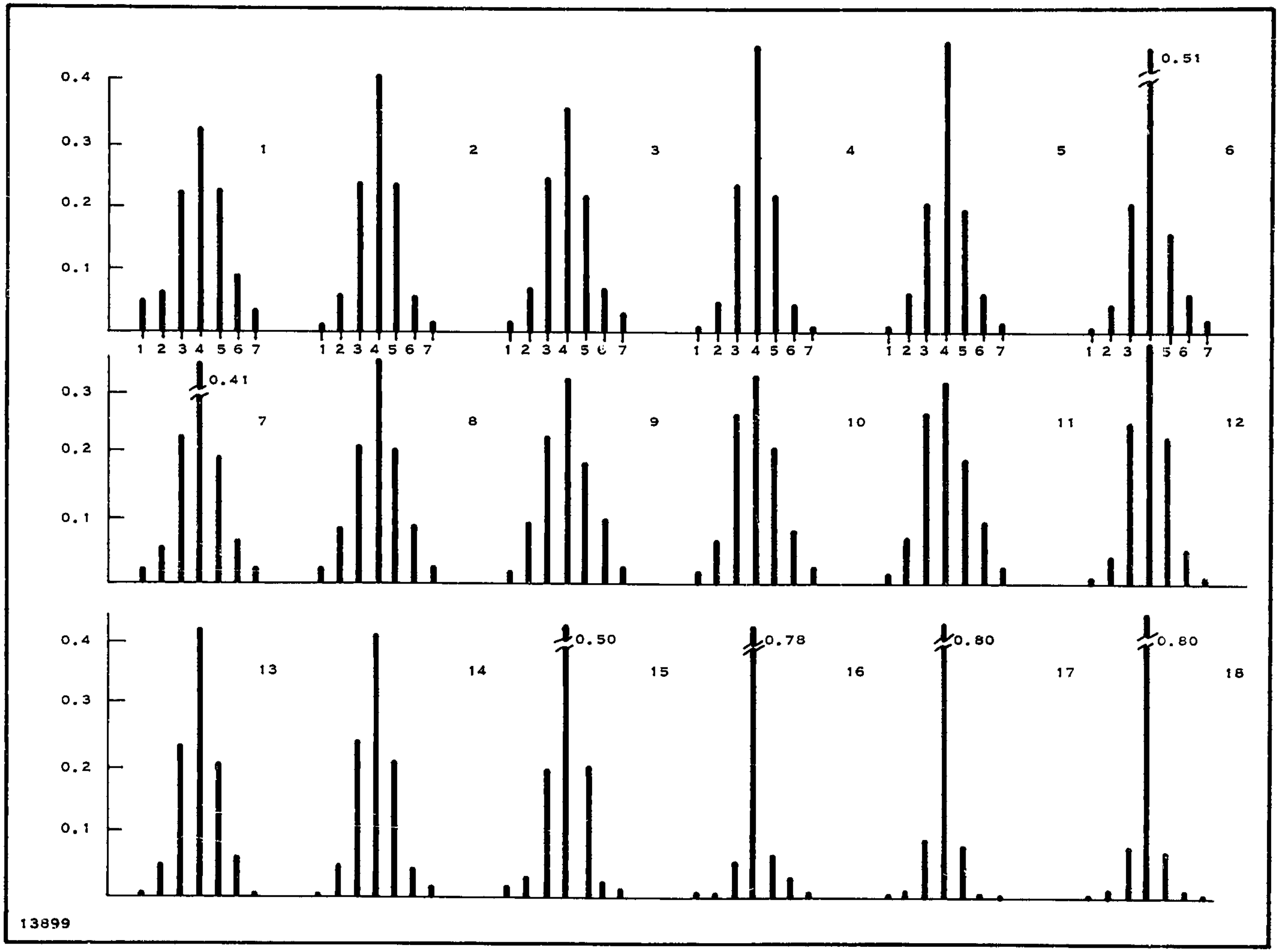


Figure 18. Distribution of Transformed Channel Values–T0101 P06A–Nonsilence Frames

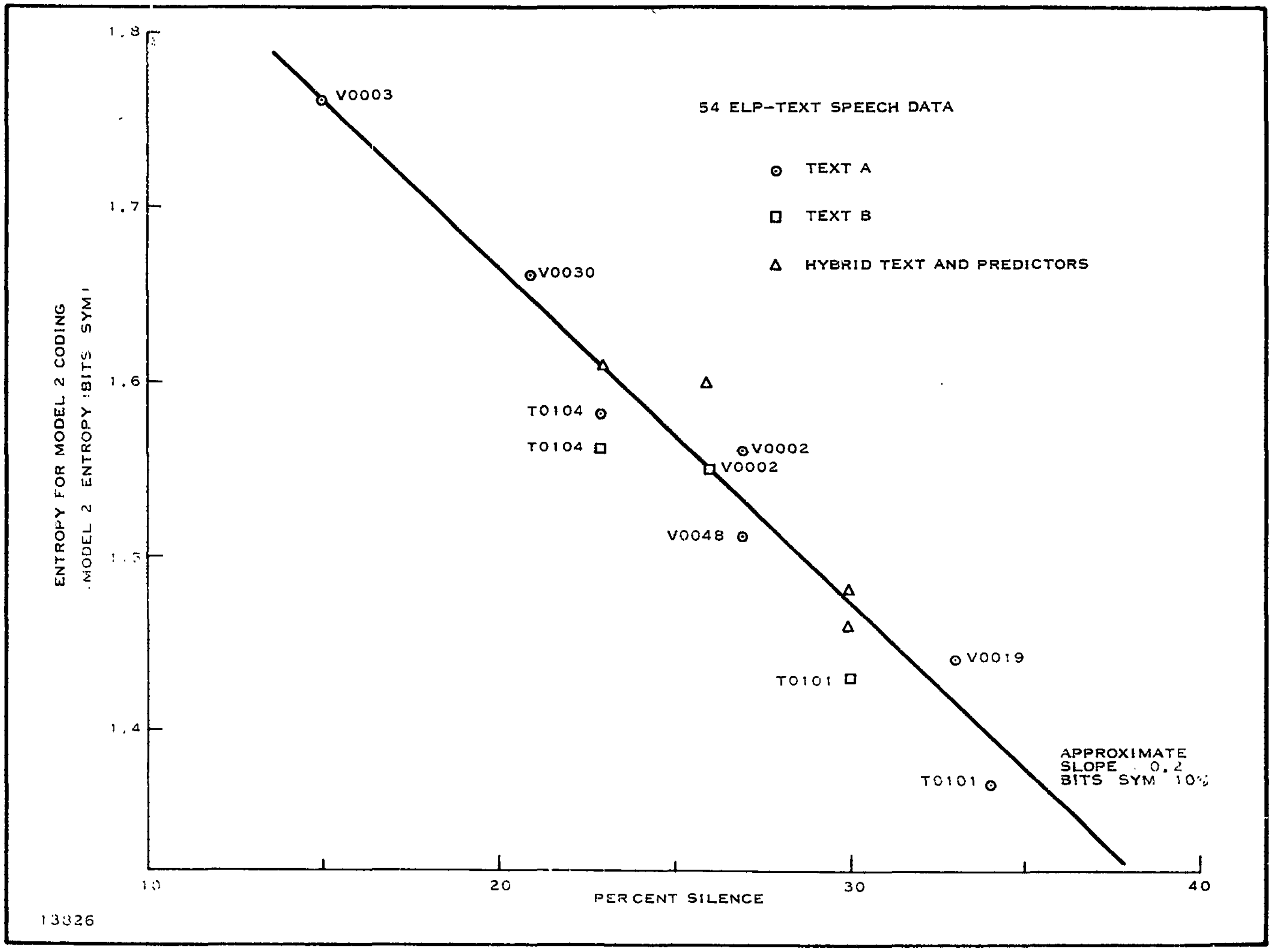


Figure 19. Model 2 Combined Channel Average Entropy

B. SIMULATION RESULTS OF HUFFMAN CODER BUFFER AND DELAY ANALYSIS

The computer simulation program to provide data for analysis of coder buffer and delay requirements was described in Section IV. Very little data was processed for this analysis because of the dependence upon type of speech sample. Each text sample provided around a 100 speech/silence cycles, and the summary of the data analysis is very much controlled by the percent and the maximum length of silence intervals. The tables in Section IV provide an illustration of how the data was summarized, and the worst-case message buffer and delay requirements.

The form of the estimates described in Section IV for Model 2 processing implied that the buffer and delay are related by

$$d = \frac{B}{m} , \tag{60}$$

where d is the total delay in coding and decoding in seconds, B is the "adequate" buffer estimate in bits, and m is the message average transmission rate in bits per second. For a 50-frame-per-second sampling rate, this equation may be written

$$d = \frac{B}{900\,r} , \tag{61}$$

where r is the message average bit/symbol rate. The message "adequate" buffer size indicated by the summary of Figure 13 would imply a coding-decoding delay of approximately 2. 27 seconds, obviously unsatisfactory for some applications. Note, however, that this delay and the associated (approx. 3000 bits) buffer are the estimates for Model 2 processing of the sample most frequently illustrated in the simulation results.

The observed worst-case requirements for essentially real-time burst coding* would require a delay of about 0. 085 second with an adequate buffer size of 160 bits. These values are more typical of general variable length coding requirements, where as the larger values are a direct consequence of smoothing the speech burst data over the silence intervals to obtain the lower average bit rates.

For any of the processing models, a typical bit/symbol measure is available (allowing for efficiency of variable length code assignment) from the entropy estimation or simulation procedures; the delay-buffer requirements are in general related linearly similar to Equation (61).

It may further be observed that whereas large percent silence implies low Model 2 bit rates, this same silence index may spell doom for the low bit rate efficiency when the associated time delays are not admissable. Of course, the speech burst bit rate is an upper bound in the sense that one could

*The speech is not spread over the silence intervals.

always transmit at this rate–not utilizing the silence to send at a lower rate. In this instance the lower buffer/delay estimates mentioned above are applicable. The adaptability of the Model 2 processing scheme to meet various system requirements is a primary virtue.

C. DISCUSSION OF SIMULATION RESULTS AND SUMMARY CONCLUSIONS

The data presented in the major portion of the report and the tabulation of data in this section are summarized by the following conclusions.

(1) The implications of early analysis and unequal length predictors has been verified by the simulation results presented for comparison in rows 3 and 4 of Table VII and Table VIII. The major, if not total, bit rate reduction through linear time invariant predictive coding is achieved with prediction no more than 2 of 3 frames across channel and back in time. Further, the accuracy of the entropy estimates computed from source statistics and the predictor algorithm have been consistantly verified for all prediction models. Table IX illustrates the degree of approximation. This latter observation should apply toward similar analysis of other type source data.

(2) The Model 1 processing scheme has been demonstrated for matched, hybrid, and difference data simulations to achieve a compression factor of 0.6 to 0.65. This model was not pursued extensively because of the overlapping implications of Model 2, and the restriction upon extrapolation of results to speech sources of other than text characteristics. Model 1 data is tabulated in columns 7 and 8 of Table VII.

(3) The most extensive predictive coding (linear time invariant) analysis has been directed toward the speech burst information rates and influencing effects. This analysis is applicable to the Model 2 processing scheme (which provides the lower bounds on bit rates achieved with simulation) and is thought to be the most readily extendable to other types of source data if delay/buffer requirements are admissible. Figure 19 provides empirical evidence of the "control of silence" of Model 2 bit rates, independent of speaker classification. Columns 15 through 18, 21 and 22 of Table VII contain the Model 2 statistics.

(4) The uniformity of simulation results for processing with the four hybrid data samples is a significant verification of the assumed ensemble properties of stationarity and ergodicity. These simulation statistics are represented by rows (3, 6), (7, 9), (10, 12) and (13, 14) of Table VII. The predictive coding transformation remained stable with hybrid simulations based upon the subjective classification of "most different" speakers. The entropy statistics for these simulations are tabulated in rows 16 and 18 for comparison with the separate samples, rows 15 and 17, and all other for that matter. The quantization noise and clipping statistics for these "worst case" simulations were comparable with all others with the single exception of additional clipping in tne simulation of row 18. This excessive clipping in one channel was obviously due to the two extremes in "high" (V0019) and "low"

(T0101) voice classifications, and occured with about 8.5 percent of the channel values in one channel.

(5) The simulation results clearly provide sufficient data for a comparison of cost versus efficiency for variable length coding with one or multiple channel coders. It appears that the savings with multiple coders could hardly be worth the expense and complexity compared with the efficiency and uniformity of the combined channel statistics for any given model.

(6) The only major compilation of differenced data statistics are presented with the speech burst statistics for one speaker in Table VIII. Differencing is by far the simplest, most uniform over speakers and models, and least expensive manner of extracting the mean symbol values. This operation alone is shown to significantly reduce the entropy, with somewhat more "clipping" than a prediction scheme yielding the same entropy. Results of differenced data simulation similar to those presented were observed for a few other speech samples with scattered length and level variations. Figure 14c is a vivid illustration of the effect of differencing without prediction. However, because the tails of the distribution do not decrease as is the case with statistical prediction, it is indicative that this scheme requires psychoacoustic evaluation more than any other. It is possible that an evaluation of this type has been performed with a "delta modulation" scheme at some other laboratory. The similarity of processing is obvious, but results of experimentation, if any, of this type with vocoded speech are not known.

(7) The other intent of compiling the data of Table VIII is to show typical entropy and clipping statistics versus predictor length and number of quantization levels. Note what appears to be a threshold of instability in the processing of five level quantized differenced data. A similar result was observed with seven-level simulation with another speaker, but not in time to test stability of either samples at lower levels of quantization. As was discussed in the fidelity considerations of Section II, excessive clipping can result in randomness or instability of the predictions, and a total loss of system fidelity.

In general, the other variations with length and quantization levels are obvious. The reduction in entropy with increased predictor length is small. The clipping excursions are reduced considerably by increasing the number of quantization levels with only a small increase in entropy. Note again that the statistical prediction scheme was stable at all ranges of quantization investigated.

(8) The histogram representation of raw and transformed symbol data provides visual evidence of the "peaking" as a result of prediction. The difficult to assess fidelity considerations discussed in Section II should be evaluated with this effect for reference. Note that reduction of the raw data of Figure 14 a to seven-level quantization would result in about a 3.5 percent "clipping" penalty; but that the seven-level prediction error data of Table VIII

suffered only a fraction of 1 percent clipping as the worst case. The variation of fidelity and entropy with predictor length and number of quantization levels may only be stated conclusively through psychoacoustic evaluation.

(9) The theoretical and simulation data presented for the linear time-invariant predictor models investigated in this study show an exceedingly high degree of consistancy. In this respect the approach is very suggestive of practical application.

The insight provided by this investigation may help to explain the limited speech burst compressions and is a basis for recommended future study which is more fully discussed below. The simulation results for the various speech burst models consistantly gave a compression factor of about 66 percent-plus or minus a few hundredths of a bit. These results suggest a rather obstinate lower bound for the fundamental approaches investigated. The compression factors approaching one-half for Model 2 processing must be appraised in conjunction with the not to be denied buffer and (especially) delay effects on a total system.

(10) The major dissapointment, and subsequential gap in the investigative study, was the unsuccessful attempt to alter the basic digital computer program to simulate the voiced/unvoiced option of Model 4. Although the options for this model and the statistical justification for its conception were discussed earlier, the multitude of other programming and data handling demands of the study prohibited complete development of this option. However, because the validity of the entropy measures estimated from theoretically minimum variance statistics has been adequately verified, the following estimates are presented to partially fill the gap.

For speech sample 6V0002P06A the ULP and ELP estimates of voiced entropy were 32. 35 and 32. 325 bits/frame respectively. The unvoiced estimates were 26. 62 and 25. 376 bits/frame respectively. For this sample there were 982 voiced frames and 245 unvoiced frames for a total of 1227 speech burst frames. Summing the properly weighted entropy measures for the sample gives 1. 73 bits/symbol for the ULP and 1. 72 bits/symbol for the ELP. Note that these estimates would compare with the "Avg. Chan H" values given in Table VII.

The most recent check-out run with the Model 4 simulation looks correct, but the trial solution has not been run. The recent run measured the "Avg. Chan H" from 54 ELP, seven-level simulation with 4 T0101P064 data to be 1. 85. The "Comb. Chan H" for the sample was 2. 00, both values in bits/symbol. The quantization noise and other statistics appear equivalent to the usual seven-level 54 ELP results.

D. RECOMMENDATIONS FOR FUTURE WORK

A number of questions which appear deserving of further study are discussed in this paragraph. These questions concern both additional results needed to apply the coding techniques discussed above in practical systems and several possible extensions of these techniques into more sophisticated methods.

1. Direct Refinements

First we shall discuss several points having to do with the development of our basic techniques for applications. Generally, these considerations are fairly straightforward and only short comments are offered.

Even if one intends to use one of the prediction schemes without modifications, it probably goes without saying that the whole relevant analysis should be repeated using speech samples which are as nearly typical of the speech to be encountered in normal operation. In this connection, careful attention should be directed to obtaining a realistic speech-to-silence ratio when speech samples are being acquired for design purposes. Also, it is recommended that in recording this sample data a method permitting accurate reproduction be used if possible. Although the effects of quantization noise were not distinguished separate of other system parameters, it is still considered to be an important effect.

The predictors analyzed have each been designed from statistics taken from a single talker's voice. Consequently, the results are actually of direct applicability only to systems which are "adaptive" in some way to individual voices, or voice classifications. Since a relatively small number of parameters is required to characterize a specific code, it would be a simple matter to have system users insert their particular parameters into the system before a conversation is begun. Of course, this may prove to be an undesirable operational feature, in which case the possibility of a genuinely adaptive system can be entertained. While it does not appear feasible to incorporate the entire measurement and synthesis procedure used in this study in an operational system, there is a very good chance that a fairly simple procedure can be made to alter a standard initial predictor to suit a given talker's voice. Probably, general procedures suited to this purpose will be reported on in literature in the near future.

A completely straightforward system in which the same predictor is used for all voices would eliminate this concern. Although we did not directly synthesize and evaluate such a system, the results that were obtained by processing one talker's voice with another talker's predictors and by other scrambled combinations seem to indicate a surprisingly high degree of uniformity between talkers from a coding viewpoint. Tentatively, it appears that fixed nonadaptive coding may be a much more valuable method that originally expected. Consequently, a direct evaluation of this possibility should be made before adaptive schemes are resorted to.

In applications where implimentation economy is a major concern, several possibilities for simplification of the basic models should be investigated. In this study no attempt was made to single out particular points as being more or less useful than others for prediction purposes. Although it is very dangerous to try to make conclusions of this type, there is some indication that a few points "contribute" most to the prediction, and all others have roughly equivalent weight. At any rate, it does seem that a deliberate search for the optimum selection of prediction points should be undertaken if equipment cost is a vital concern. As an alternate and perhaps more easily exploited approach, one can use the basic techniques to describe earlier to look for useful arrangements of a smaller number of predictors than the 18 used uniformly in this study.

In those codes where the VAP or V/UV items are used for logical switching, a problem arises concerning the transient behaviour of the prediction loop. To minimize the gyrations of the error signal when a predictor is first "turned on", the shift register should be initially loaded with appropriate values. In our simulations a simple scheme was quickly found which adequately served the requirements of the study. It is not necessarily the best possible scheme, and some effort should be directed to the problem in any developmental program.

No careful consideration has been given to the effects of channel errors which are inevitable in practical systems. Ordinarily these errors pose a severe problem in predictive coding; however, in the case of speech, the problem should be less troublesome, since through appropriate logic the frequent occurence of silent intervals can be utilized to restore the receiver shift register to initialization contents. Whether or not such a provision is adequate should be checked.

2. Alleviation of the System Delay Problem

As was noted earlier, spreading speech burst data over ensuing silent intervals requires that the speech message suffer a net delay in passing through the transmitter and receiver buffers. As has also been noted, this delay can easily approach the point where it is objectionable to conversing parties. In a very real sense the tradeoff between average bit rate and net delay is fundamental and unavoidable as long as one considers simple duplex systems. If the total system delay is to be kept small, burst data must be transmitted at a relatively high rate. If this is done, then the silent intervals are wasted in so far as the speech communications are concerned. Clearly if a channel is to be utilized efficiently when low-delay coding is employed, then the silent intervals must be used to transmit data from some second information source. In Part III of the report prepared on this contract[12], the use of silent intervals for teletype data transmission is discussed.* Another possibility for efficiently utilizing the silent intervals should also be investigated. If the channel in question is not a single voice channel but is a higher capacity channel servicing several conversations, then straightforward interleaving of speech burst data might lead to efficient channel utilization without severe delays in the individual conversations. Whether

*Results of this study disclosed that as many as 38, 75 wpm teleprinters can be multiplexed into the silence frames of a full duplex vocoded speech system.

or not such a system is practical, meets a need, and solves the problem should be investigated.

3. Integration of the Quantization and Coding Functions

Throughout this report a major concern has been the possible effects on system fidelity resulting from relocating the quantization process. On the whole, we have tried to keep the net or effective quantization process as nearly unchanged as possible, so that the compression achieved could be attributed to the coding process per se without qualifications. Of course, the significance of the clipping effect discussed earlier still needs to be determined subjectively. Above and beyond this concern however, there are several other questions and possibilities associated with the relocated quantization process; all in all it seems that a new "round" of subjective experiments is in order.

Assuming that the original purpose of leaving system fidelity unchanged has been accomplished, we can now broaden our deliberations. In particular it can be asked whether or not the quantization procedure most appropriate for the original input signals is actually the best one to apply on the error signals resulting from prediction. In particular, the merits of nonuniform quantization interval widths should be reinvestigated; experience with other signals, and video signals in particular, has shown that a nonuniform quantization of error signals frequently yields the best subjectively determined fidelity, even though uniform quantization is best in the absence of prediction. Generally one finds that the size of the intervals should increase with the magnitude of the error signal.

Besides the possibility of an improvement in fidelity, there is another motivation for considering a modification of this type to the original quantization scheme. A procedure which gathers levels near the zero-error point and spreads levels for large error values tends to equalize (up to a point) the occurence probabilities of the various error symbols. In particular, such a scheme can be deliberately designed to make the probability of an error signal value falling in a given interval the same for all intervals. This of course eliminates the need for variable length code assignments and may have advantages in applications where cost is a paramount concern.

The rather unattractive buffer sizes needed to spread speech burst information over silent intervals suggests another modification to the quantization procedure. Probably, it is the relatively infrequent occurence of several long speech bursts combined with short silent intervals which determines the memory sizes found in the study. To cut down on these peak loads imposed on the buffer system, one might gradually increase the quantizer interval widths as the buffer fills up. In this way a controlled fidelity degradation can be substituted for the catastrophic breakdown otherwise associated with buffer saturation. In such a scheme, the fidelity of reproduction for each input symbol is established at the instant it is quantized; this implies that the control system must act without specific knowledge of the durations of following speech and silent intervals. However, because the data, after coding, usually sits idly in the buffer for some time before it is actually transmitted, one could place a part of the buffer ahead of the coder-quantizer combinations, so that the latter is operating on a backlog of input data. A configuration such as this probably could achieve much better regulation of a variable-fidelity quantization process. If indeed peak loads arrive at the buffer very infrequently, then perhaps the overall system delay could be substantially reduced in exchange for infrequent temporary degradations of fidelity.

Although it does not particularly imply a need for further studies, there is another reason for considering variable-fidelity quantization. Different talkers will generally have different average entropies measured with respect to a standard fidelity criterion. For adjusting each talker's bit rate to that of a standard fixed rate channel, expansion or contraction of the quantization intervals could be used.

4. Extension of the Linear Predictor Model

Although simplicity of implementation is the only advantage of linear prediction methods explicitly discussed in the forgoing, there are of course more positive reasons for having chosen this basic model for initial study. These other motiviations generally derive from considerations of the usual source-filter model of the vocal apparatus and the nature of the decomposition of speech waveforms affected by vocoders. These same considerations, however, also serve to point out the inherent limitations of the relatively simple code models studies to date and to suggest the embellishments one should add to these basic models in order to obtain greater compression factors.

When one examines sufficiently short segments of a sonogram (on the order of 0.1 to 0.2 second, say) it is immediately apparent that there is a high degree of "organization" in the source process. Of course, when larger segments are examined, it is also apparent that this "organization" is rather short-lived; that is to say, the exact nature of the source process clearly changes considerably over successive short intervals of time. If the term has any meaning at all, one can certainly call vocoded spectrum data a "quasi-stationary" process.

The simplest predictor mentioned in this report is one in which the same predictor is used on the entire multiplexed input data stream. This scheme was of course abandoned immediately after its mention in recognition of the fact that the statistics of the multiplexed sequence changed with time–specifically the change cyclicly with a period of 18 symbols. Next, the differencing feature was substituted for subtraction of a fixed mean value largely in recognition of the fact that means, figured on a short term basis, changed with time. Similarly the use of predictors operating on nonsilence data only and the use of separate predictors for voiced and unvoiced data both represent attempts to capitalize on more of the "quasi-stationary" characteristics contraining the source process. Certainly the results have shown the merit of even these particularly simple steps.

If the redundancy associated with short term structure is to be removed, then the coding process itself must change rapidly in accordance with the changes in the source process. Obviously the coders we have analyzed do not have this sort of capability except in the simplest of senses; they are sensitive to, or effective against, only that redundancy apparent in statistics obtained through averages taken over very long time intervals. To obtain more compression from linear predictors, means must be developed to introduce more short-term time variations of the predictors in direct concert with those of the source.

Useful extensions of this type must come about largely through sheer inventiveness because theoretical approaches certainly will be of little help in the conceptional stages. The only readily apparent sources of guidance are the various established models of the speech production process. For the present at least, it seems that any sensible approach to time varying predictors must depend heavily upon the source descriptions provided by these models. In fact, it may even be that the predictors should directly implement one or more of the operations contained in the various models.

At one extreme one can start with the simple "first-order" phonemic model, which describes speech as a sequence of discrete, disjoint, and largely independent sounds following one after the other in time. Corresponding to this model one can consider a coding system incorporating a battery or library of predictors in which each one is tailored to handle a particular sound or limited class of sounds, and from, which, individual predictors could be selected for use depending upon the particular speech sound present at the input. On the other hand, even though the motivation behind this sort of operation derives from a drastically oversimplified source model, it certainly seems reasonable that something at least roughly resembling a phonemic breakdown would be appropriate if very large compression is to be achieved. Since the motivation for this system would seem to imply that very complicated decision rules must be used, we should point out that it is by no means essential to have the selection decisions resemble a "conventional" segmentation or phonemic classification in any way whatsoever. The decision rules can be made very simple and a potentially useful system is still possible. Such a system is really a straightforward extension of the voicing option scheme analyzed above, except that now the selection decisions would be based upon the spectrum data itself rather than on nonspectrum data.

To illustrate this type of system with a very simple case, suppose the entire bank of predictors were operated simulataneously. To determine which predictor output to use in coding a given symbol or frame, the error signals themselves could be examined to see which predictor was giving the best results at that time. In some arrangements it would be necessary to send "keying" information to the receiver to indicate which predictor was used. On the other hand, if only the error signals generated on past frames were used in the selection decision, then the decision could be duplicated at the receiver and no auxilliary keying information need be transmitted.

Thus, except for a little logic, this particular system is made more complicated than those analyzed in our study only through the additional predictors required.

The substantial shortcomings of simple phonemic models in describing connected speech do serve to thoroughly complicate the design and evaluation of multipredictor schemes. Certainly at this time one can only engage in speculation as to the probable effectiveness of such schemes. In this connection, however, it should be pointed out that the spectrum pattern data now being collected and analyzed at AFCRL should be of direct and inestimable value to any future studies on this subject. If a segmentation and classification of speech sounds can be affected using this data and technique, then at least a first attempt at synthesizing a multipredictor system can be made using the same method employed in the present study.

At the other extreme are the source models which describe speech in terms of its formant structure. The "organization" of speech spectrum data into a few slowly migrating formants unquestionably represents the most important single source constraint from a general predictive coding view-point. Presumably its existance implies the existence of considerable redundancy in the vocoder representation of spectrum information. On the other hand, it is very easy to convience one's self that the linear, time invarient, predictors studied are probably very ineffective in capitalizing on this source constraint. While formant structure induces a strong short-term correlation between spectrum symbols in the general sense of correlation, only a weak or almost "accidental" correlation exists in the specific linear sense which is relevant to linear prediction. Although the multipredictor schemes suggested above might be more effective, they still do not take formant structures explicitly into account and hence do not represent a direct assualt on the problem of making better use of this characteristic of speech.

Possibly a scheme can be devised whereby the formant frequencies can be estimated explicitly and then used to determine the prediction coefficients to be used. In so far as the formant frequencies themselves are concerned, the estimation problem is surmountable; and, once estimated, this "control" information could be handled much like the selection decisions in multipredictor systems. That is it could either be produced by duplicate processes at transmitter and receiver or it could be transmitted through the channel as auxilliary information. Determination of the predictor coefficients from these signals is of course the central problem and it appears to be a problem of substantial magnitude at the very least. In this case, rather than just selecting a predictor from a library, one presumably must invent a means for generating the predictor through some computational procedure. Since no known practical synthesis procedures are available for such a system, exploitation of this possibility probably would require a substantial investigation before any one approach is envisioned and an attack formulated.

It seems apparent from the limitations of methods analyzed in this study and recent formant tracking results that substantially greater compression factors could be obtained by resorting to sufficiently sophisticated short term time-varying linear predictors. While the implementation of such systems is not necessarily difficult, the associated analysis and design problems certainly are difficult.

SECTION VI

REFERENCES

[1]C. E. Shannon, "A Mathematical Theory of Communications," Bell Systems Tech. Journal (1948).

[2]N. Wiener, "Extrapolation, Interpolation, and Smoothing of Stationary Time Series," Wiley (1949).

[3]P. Elias, "Predictive Coding," IRE Transactions on Information Theory" (March 1955).

[4]A. N. Kolmogoroff, "Interpolation and Extrapolation," Bulletin de l'academie des Sciences de USSR, Ser. Math., 5 (1941).

[5]Fant, Gaunnar, "Acoustic Theory of Speech Production," Mouton and Co. (1960).

[6]C. P. Smith, "Normalization of the Voice Spectrum" (Abstract), J. Acous. Soc. Am., Vol. 29 (1957), p.777.

[7]B.A. Morris, "An Escalator Process for the Solution of Linear Simultaneous Equations," Phil. Mag., Vol. 37 (Ser. 7) (1946).

[8]Texas Instruments Incorporated, "Speaker Recognition," Final Report, Part I, Contract AF 19(628)-345 (May 1963).

[9]D. A. Huffman, "A Method for the Construction of Minimum Redundancy Codes," Proc. IRE, Vol. 40 (September 1952), p. 1098.

[10]P. G. Neumann, "On a Class of Efficient Error-Limiting Variable Length Codes," Bell Telephone Laboratories Report.

[11]E. N. Gilbert and E. F. Moore, "Variable Length Binary Encodings," Bell Systems Tech. Journal (July 1959), p.933.

[12]Texas Instruments Incorporated, "A Study of the Feasibility of Multiplexing Teletype Data into Nonspeech Time of Vocoded Speech Transmission," Final Report, Part III, Contract AF 19(628)-345 (June 1963).

SECTION VII

SCIENTISTS CONTRIBUTING TO REPORT

Persons contributing to this report were Mr. R.L. Brueck, Mr. A.R. Aitken, and Mr. D.R. Ziemer. A brief resume of the education and experience of each follows.

BRUECK, R.L.

M.S. in Applied Mathematics, University of Colorado
B.S. in Applied Mathematics, University of Colorado
Member of Society of Industrial and Applied Mathematics, Mathematical Association of America, Tau Beta Pi, and Sigma Pi Sigma
Author of contributed papers to the Society of Industrial and Applied Mathematics and The Acoustical Society of America

Mr. Brueck is currently conducting an investigation of predictive encoding techniques to speech bandwidth compression. Prior to this, he was engaged in theoretical analysis of optical correlator techniques for A/J communication application. He joined Texas Instruments in 1962 after three years on the teaching staff as Instructor in the Department of Applied Mathematics of the University of Colorado in Boulder. Within the Department of Applied Mathematics, his primary teaching and research responsibility was in conjunction with the Numerical Analysis Center, assisting various university research efforts. While teaching at the University of Colorado, he also was Consultant for the Engineering Experiment Station, a part-time employee of Ball Brothers Research Corporation in Boulder, and a participant in NSF-sponsored research program (Research Associate). Mr. Brueck was awarded a stipend for participation in MSF Numerical Analysis Summer Institute, UCLA, in the summer of 1960. During the summer of 1959, he was staff mathematician with Shell Oil Company in Denver. His areas of interest in graduate study include mathematical physics, theory of nonlinear differential equations, spectral theory of linear operators, and numerical analysis.

AITKEN, A.R.

B.S. in Geophysics, Massachusetts Institute of Technology
Patent Granted 2 May 1962 -- Digital Seismic Exploration System

Mr. Aitken joined Texas Instruments in 1956 and is presently working on the application of predictive encoding to the problem of reducing channel requirements for speech communication. Prior to this, he worked on applications of statistical decision theory to detection systems, and on the development of an airborne multiple maser magnetometer detection system. For several years at Texas Instruments, Mr. Aitken has been responsible for developing acoustic and magnetic methods of submarine detection, signal detection theory, and digital data handling including digital-analog converters, digital computer circuitry, and logical design of various special-purpose computers.

ZIEMER, D.R.

M.S. in Engineering, University of Texas
B. S. in Engineering, Iowa State University
Member of IRE and Eta Kappa Nu

Mr. Ziemer joined Texas Instruments in 1957 and currently is in charge of a group studying advanced communication systems concepts. Previously, he was in charge of a group studying the problems associated with speech bandwidth compression techniques as they apply to digital voice communication systems. Prior to this assignment, Mr. Ziemer was engaged in investigating imagery processing, interpretation, and bandwidth compression techniques. His previous experience includes one year as senior engineer on a missile program for Temco Aircraft Corporation, and from 1953 to 1956 he was a research engineer with the University of Texas Electrical Engineering Laboratory, being primarily concerned with system analyses and data reduction. While at the University of Texas, Mr. Ziemer also taught in the Electrical Engineering Department. Other experience includes two years with Chance Vought Aircraft Corporation, where he was concerned with the analysis of automatic control systems.

AD –

Texas Instruments Incorporated, Dallas, Texas
EFFICIENT UTILIZATION OF CHANNEL CAPACITY FOR SPEECH COMMUNICATION, by R. L. Brueck, A.R. Aitken, et. al., July 1963. 96 p. incl. illus. (Proj. 4610: Task 461002) (Report 15-73801-15; AFCRL-63-316)
Contract AF 19(628)-345 Unclassified Report

This report describes a research investigation directed toward more efficient utilization of channel capacity for speech communication. This objective was pursued by a program (1) to theoretically analyze the benefits realized by "predictive encoding" of voiced speech sources, (2) to propose and analytically design a model of a typical processing system utilizing predictive coding, and (3) to evaluate the performance characteristic of such a

over

UNCLASSIFIED

1. Speech transmission
   – Mathematical
   analysis

I. Brueck, R. L.

UNCLASSIFIED

system by simulation with vocoded speech samples on a digital computer. The results were significant in that a compression of 35 to 40 percent was obtained relative to the initial requirements for transmission of the spectrum protion of vocoded speech data. This magnitude of compression can be obtained for essentially real-time transmission and buffer storage requirements of the order of 100 bits. Compression factors greater than one-half are possible if a time delay in the speech transmission can be tolerated and if additional memory can be supplied.

AD –

Texas Instruments Incorporated, Dallas, Texas
EFFICIENT UTILIZATION OF CHANNEL CAPACITY FOR SPEECH COMMUNICATION, by R. L. Brueck, A. R. Aitken, et. al., July 1963. 96 p. incl. illus. (Proj. 4610: Task 461002) (Report 15-73801-15; AFCRL-63-316)
Contract AF 19(628)-345 Unclassified Report

This report describes a research investigation directed toward more efficient utilization of channel capacity for speech communication. This objective was pursued by a program (1) to theoretically analyze the benefits realized by "predictive encoding" of voiced speech sources, (2) to propose and analytically design a model of a typical processing system utilizing predictive coding, and (3) to evaluate the performance characteristic of such a

over

UNCLASSIFIED

1. Speech transmission
   – Mathematical
   analysis

I. Brueck, R. L.

UNCLASSIFIED

system by simulation with vocoded speech samples on a digital computer. The results were significant in that a compression of 35 to 40 percent was obtained relative to the initial requirements for transmission of the spectrum protion of vocoded speech data. This magnitude of compression can be obtained for essentially real-time transmission and buffer storage requirements of the order of 100 bits. Compression factors greater than one-half are possible if a time delay in the speech transmission can be tolerated and if additional memory can be supplied.